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, 23666, USA
[2] NASA Langley Research Centre, Hampton, 23666, USA

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

Abstract. This study applies Fuzzy K-Means (FKM) cluster analyses to a subset of the parameters reported in the CALIPSO lidar level 2 data products ~~and compares the clustering results with~~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 CALIPSO data products. ~~The selection of samples~~FKM is an unsupervised learning algorithm, whereas the CALIPSO 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 ~~parameter analyses in feature type discrimination are discussed. Statistical results show that the FKM classification agrees with the CAD algorithm classification for more than 94% of the cases in troposphere. This is because the lidar measured signatures of most clouds and aerosols are naturally different. Based on their different natures, objective methods can effectively separate clouds from aerosols in most cases. In addition to validating the current CAD algorithm, the FKM clustering can also provide new insights and supplemental information to help better understand the driving parameters in the scene classification process.~~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 CALIPSO 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 Observations (CALIPSO) mission has been developed through a close and on-going collaboration between NASA Langley Research Center (LaRC) and the French space agency, Centre National D'Etudes Spatial (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 and Vane, 2007). The availability of continuous, vertically resolved measurements of the Earth's atmosphere at 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), on-board CALIPSO, is the first satellite-borne polarization-sensitive lidar that specifically measures the vertical distribution of clouds and aerosols along with their ~~microphysical and~~ optical and geometrical properties. The level 1 CALIOP data products report vertically-resolved total atmospheric backscatter intensity at both 532 nm 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~~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 km to 20.2 km, and 180 m from 20.2 km 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 km, 1 km and 5 km). The layer products consist of a sequence of column descriptors (e.g., latitude, longitude, time, etc.) 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 Algorithms (SCA), and the Hybrid Extinction Retrieval Algorithms (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 ~~(Vaughan et al.,~~. 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 sub-modules: the cloud and aerosol discrimination (CAD) algorithm (Liu et al., 2004, 2009, 2018), 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., 2018). 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.

~~The CAD algorithm uses~~ 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 ~~two classes, cloud~~clouds and ~~aerosol~~aerosols (Liu et al., 2004, 2009, 2018). ~~The CAD algorithm has been improved over many years, from using three independent measurements (~~Using a larger number of layer ~~mean attenuated backscatter at 532 nm, the layer integrated 1064 nm to 532 nm volume color ratio, and the mid layer altitude) in version 1 (V1) to five independent measurements (adding layer integrated 532 nm volume depolarization and latitude) in version 4.1 (V4). For the operational CAD PDF method, using more measured independent information~~ attributes (i.e., higher dimension PDFs) ~~is expected to yield~~generally yields increasingly accurate cloud and aerosol discrimination.

~~In addition to the CALIPSO team, scientists over the globe also~~ While both V3 and V4 COCA algorithms use the ~~CALIOP data and products, and their work greatly contributes to the evaluation of the CAD products and helps to better understand lidar techniques for distinguishing clouds from aerosols (Chen et al. 2010; Jin et al., 2014; Di Pierro et al. 2011). Using~~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, 2018; Vaughan et al., 2018). The V4 PDFs have been re-built to better discriminate dense dust over the Taklimakan desert, 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 resolute (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., 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 (Teche 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 on-going 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 classificaiton 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 cloud and aerosol 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 probability density functions (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, etc.). 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., 2018). 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, … X_m), separate multidimensional PDFs are constructed for clouds ($P_{cloud}(X_1, X_2, \ldots X_m)$) and aerosols ($P_{aerosol}(X_1, X_2, \ldots X_m)$). Discrimination between clouds and aerosols for previously unclassified layers is then determined using

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

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

The function f is a normalized differential probability, which value ranges from -1 to 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; Liu et al. 2018). Within the CALIOP level 2 ~~layer~~data products, a percentile (integer) value of 100 × 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. ~~The algorithm has been applied to Cloud Physics Lidar (CPL) Data from 2003 THORPEX PTOST campaign and to desert dust data acquired during the Lidar In-space Technology Experiment (LITE; Winker et al., 1996), and was shown to work well with both data sets (Liu et al., 2004).~~ Because the nature of clouds is quite different from aerosols, ~~distinguishing between the two should generally~~most cloud and aerosols can be ~~straightforward~~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 operational CAD algorithm uses manually-derived, multi-dimensional PDFs to distinguish clouds from aerosols using a statistical discrimination function. In this paper, we introduce the Fuzzy K-Means (FKM) method, an unsupervised clustering algorithm, and use it for differentiating clouds from aerosols. The purpose of the study is twofold. First, by using an unsupervised clustering algorithm that is quite different from and independent of the operational method, we can validate the~~

~~results of V4 CAD algorithm. Second, an unsupervised algorithm can help us better understand classifications made by the operational CAD algorithm. For example, the FKM approach can help determine which individual parameter is most crucial in the discrimination of clouds and aerosols. It can also evaluate the degree of improvement to be expected when adding observational dimensions to the PDFs and help estimate the biases that are coming from background noise.~~

~~A description of cluster analysis and the FKM method is given in section 2. In section 3, individual cases and statistical classification results from FKM clustering are produced and compared with classifications made by the operational V3 and V4 CAD algorithms. Error analyses are performed in the following section, including error propagation, key parameter analysis, fuzzy discriminant analysis and principle component analysis. Conclusions and perspectives are given in the last section.~~

~~2 Cluster analysis~~

~~Cluster analysis is a useful statistical tool to group data into several categories and can be 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). In our study, we apply the FKM clustering algorithm to CALIOP level 2 observations to distinguish between clouds and aerosols. CALIOP directly observes three quantities: height resolved profiles of parallel and perpendicular backscatter intensity at 532 nm and total backscatter intensity at 1064 nm. The V4 CALIOP CAD algorithm distinguishes clouds from aerosols using five parameters reported in the level 2 layer products: layer mean attenuated backscatter at 532 nm, layer integrated depolarization ratio at 532 nm, layer mean total attenuated color ratio (1064 nm / 532 nm) and layer altitudes and latitudes (Liu et al., 2004, 2009). We use four of them (excluding latitude) as our set of observations and incorporate them directly into a neural network cluster analysis.~~

~~There are many different types of neural network clustering methods, such as K-means (KM), fuzzy K-means (FKM), Expectation Maximization (EM) and k-harmonic means (Nock and Nielsen, 2006). In this paper, we focus on the FKM method because the memberships belong to [0 1], and thus are comparable to operational CAD score and more logical for the cloud-aerosol discrimination task. The FKM algorithm is one of a popular class of center-based clustering algorithms.~~ 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, $\chi' = \langle\beta'_{1064}\rangle/\langle\beta'_{532}\rangle$, and mid-layer altitude, z_{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, δ_v, and the latitude of the horizontal mid-point of the layer (Liu et al., 2018). Within the CALIPSO analysis software these PDFs are implemented as 5-D arrays that function as look-up tables.

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 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., $<\beta'_{532}>$, χ', δ_v, and z_{mid}. We make this choice because clouds and aerosols show distinct centers in the $<\beta'_{532}>$, δ_v, χ', z_{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 ~0.2 to ~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).

~~Different from the KM method, FKM clustering gives every point a degree of membership in all categories rather than belonging completely to just one category.~~ 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}, $i = 1,\ldots,n$; $j = 1,\ldots,k$) ~~all~~ are assigned values between 0 and 1 (Eq. 2). When elements of the membership matrix ~~are equal to~~ $m=1$, 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 ~~in the KM method,~~(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 one (Eq. 3), and there will be at least one individual with some non-zero membership belonging to each class (Eq. 4),4). These defining relationships are written as

$$m_{ij} \in [0,1],\ i = 1, \ldots n; j = 1, \ldots, k, \quad (2)$$

$$\sum_{j=1}^{k} m_{ij} = 1, i = 1, \ldots, n, \quad m_{ij} \in [0,1], i = 1, \ldots, n; j = 1, \ldots, k, \quad (2)$$

$$\sum_{j=1}^{k} m_{ij} = 1, i = 1, \ldots, n, \text{ and} \quad (3)$$

$$\sum_{i=1}^{n} m_{ij} > 0, j = 1, \ldots k. \quad (4)$$

$$\sum_{i=1}^{n} m_{ij} > 0, j = 1, \ldots, k. \quad (4)$$

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; Bezdek, 1984; McBratney and Moore, 1985). The functional form of *J* is

$$J(M, C) = \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{\phi} d_{ij}^{2}(x_{il},) c_{jl} \quad (5)$$

$$J(M,C) = \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{\phi} d_{ij}^{2}(x_{il}) c_{jl}, \quad (5)$$

where C (c_{jl}; j=1,.., k; l =1,..., 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 (i.e.g., the Mahalanobis distance; see sectionSect. 2.3). The objective function is the squared error from class centers weighted by the ϕ^{th} power (fuzzy weighting exponent) of the membership values. For the least meaningful value, ϕ = 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 ϕ tends to degrade memberships towards the fuzziest statefuzzier states where there are more overlaps between the boundaries of clusters. MinimizationFor a specified value of ϕ, minimization of objective function *J* providesoptimizes the solutions for the best 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 ϕ^{th} ~~function~~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 individual and the centers of all classes (Eq. 7~~); i.e.,~~), or

~~$c_{jl} = \frac{\sum_{i=1}^{n} m_{ij}^{\phi} x_{il}}{\sum_{i=1}^{n} m_{ij}^{\phi}}$ $j = 1,2, \dots k; l = 1, 2, \dots p$, and (6)~~

~~$m_{ij} = \frac{d_{ij}^{-2/(\phi-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\phi-1)}}$ $i = 1,2, \dots, n;\ j = 1, \dots, k.$ (7)~~

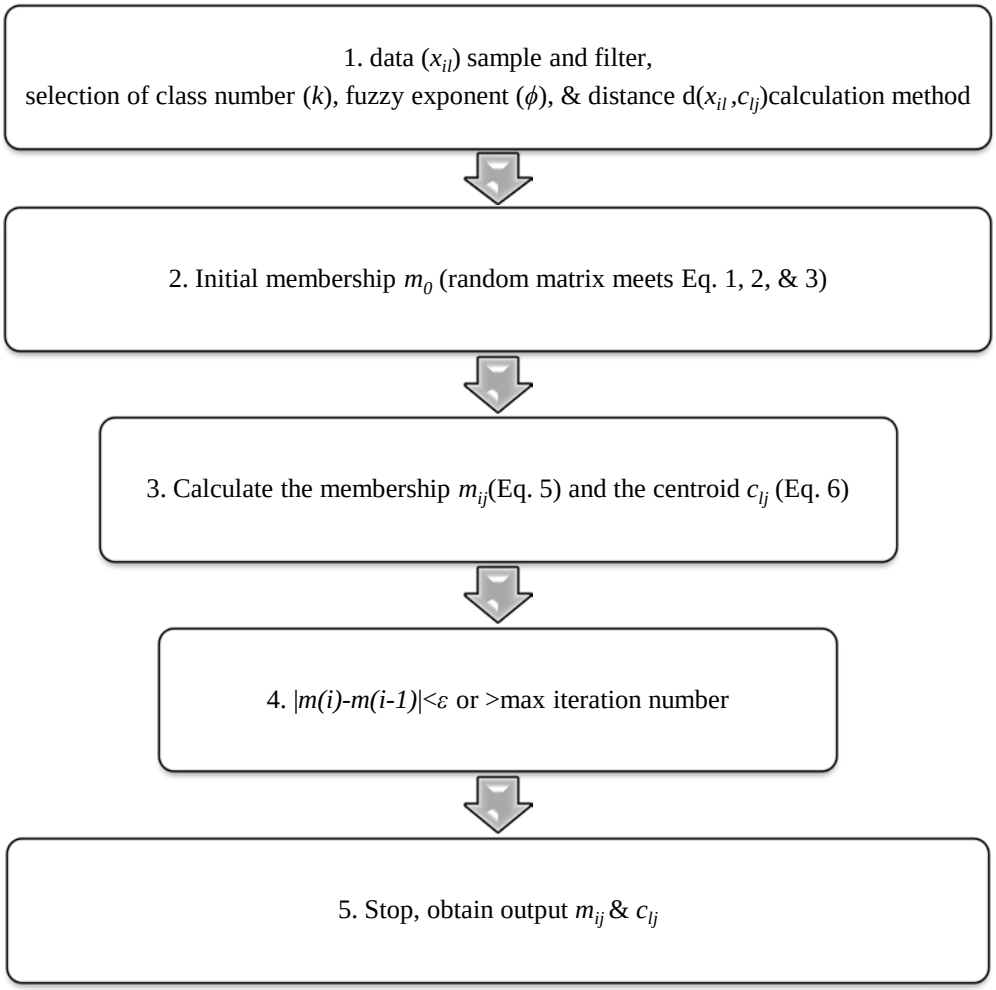
$$c_{jl} = \frac{\sum_{i=1}^{n} m_{ij}^{\phi} x_{il}}{\sum_{i=1}^{n} m_{ij}^{\phi}}, j = 1,2,\dots,k; l = 1,2,\dots,p$$

, and (6)

$$m_{jl} = \frac{d_{ij}^{-2/(\phi-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\phi-1)}}, i = 1,2,\dots,n; j = 1,2,\dots,k$$

. (7)

To obtain centroid (Eq. 6) and membership (Eq. 7) solutions, Picard iterations (Bezdek et al., 1984) are applied until the centers or memberships are constant to within some small value (~~Figure 1,~~see the algorithm flowchart in Fig. 1). We first initialize the memberships as random values using a uniform distribution that satisfies all conditions given by equations ~~1,~~ 2, 3 and ~~3. After that, we~~4. We then calculate class centers and recalculate memberships according to the new centers. If the new memberships do not change compared to the old ones (or change only within a small difference ε), the clustering process ends. Otherwise we recalculate the new centers and new memberships. ~~This process is repeated until~~If the algorithm does not converge after a fixed number of iterations, the procedure is reinitiated using newly (and again randomly) specified initial cluster centers. This process repeats until the algorithm converges to a point where the relative change in the objective function (calculated from Eq. 5, which quantifies the changes in both the memberships and centers) is less than ε (0.001~~).~~) and saves the best memberships and centers that result from the optimum random initiation corresponding to the least objective function.


1. data (x_{il}) sample and filter,
selection of class number (k), fuzzy exponent (ϕ), & distance d(x_{il},c_{lj})calculation method
2. Initial membership m_0 (random matrix meets Eq. 1, 2, & 3)
3. Calculate the membership m_{ij}(Eq. 5) and the centroid c_{lj} (Eq. 6)
4. $|m(i)-m(i-1)|<\varepsilon$ or >max iteration number
5. Stop, obtain output m_{ij} & c_{lj}

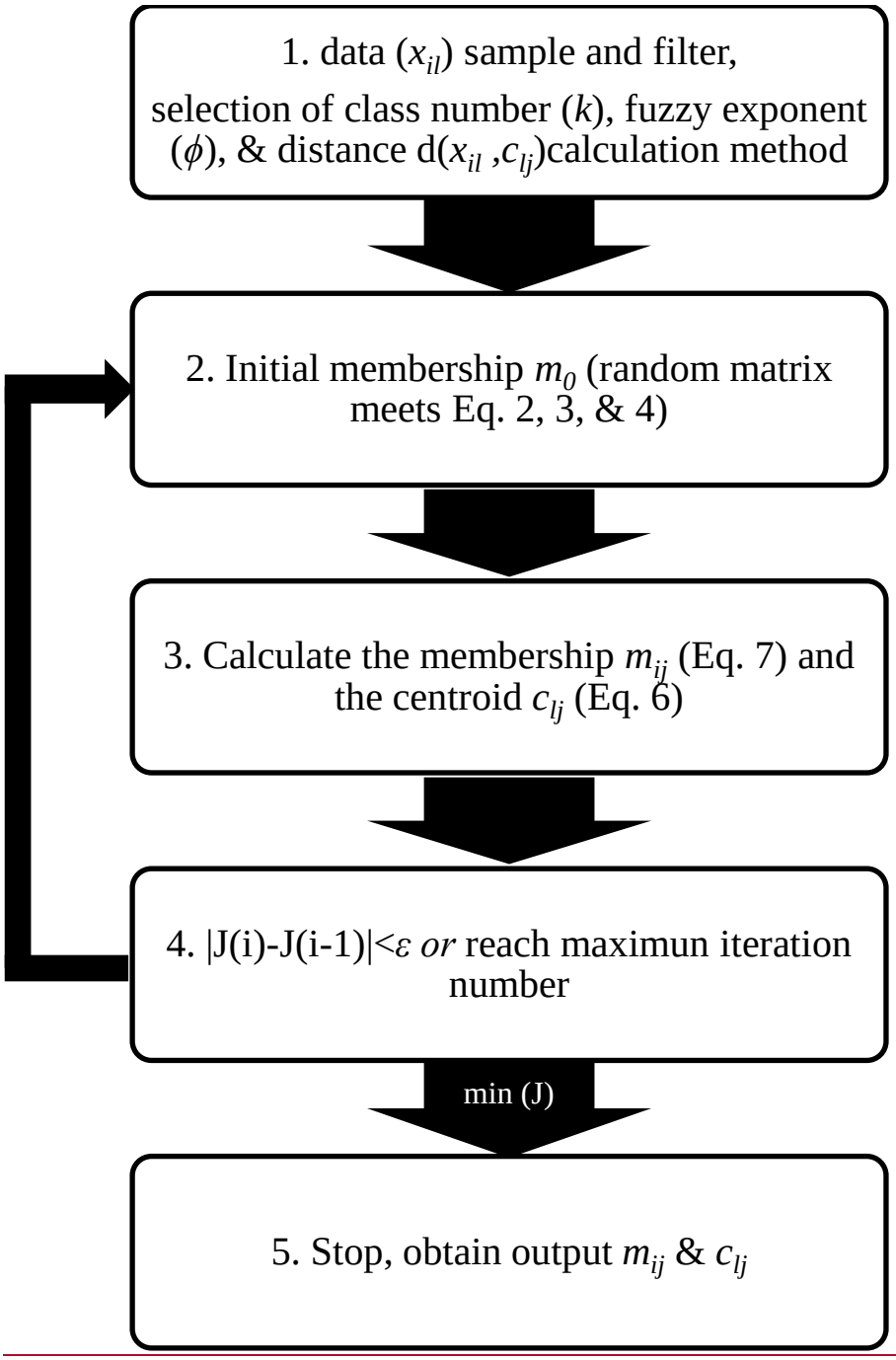


Figure 1: Flowchart ~~of~~ illustrating the operation of the Fuzzy K-means algorithm

Before running the FKM code (from Minasny and McBratney, 2002), we prepared our data by sampling, training and filtering (~~sections~~Sect. 2.1 and Sect. 2.2). We also selected ~~optimal values for class number and fuzzy exponent (section 2.4) and chose~~ a reasonable method to calculate the distance between individuals and centers (~~section~~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~~.1~~ Data sample and training

As mentioned above, four level 2 parameters are used for our cluster analysis: layer mean attenuated backscatter at 532 nm, $<\beta'_{532}>$, layer integrated volume depolarization ratio at 532 nm, δ_v, total attenuated backscatter color ratio ~~(mean 1064 nm divided by mean 532 nm), χ',~~ and mid-layer altitude, z_{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. ~~Latitude is not basic information to distinguish cloud from aerosol, but we can apply~~$<\beta'_{532}>$, δ_v, and χ' are the ~~FKM method in particular locations or seasons depending on~~fundamental lidar-derived optical properties that form the ~~purpose~~basis for our discrimination scheme. We also include altitude, as the joint distributions of ~~the study.~~ altitude with the various lidar optical properties have proved to be highly effective in identifying different feature types.

In this ~~paper,~~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 clear scenes are included, clustering could possibly provide a means of identifying different surface types. With only cloudy data, 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 data set 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 ~~select~~ ensure the selection of an appropriate training data~~, a PDF for the smaller sample that~~ set, the shapes of the PDFs of the relevant parameters derived from any proposed training set should closely ~~matches the shape of~~ match the shapes of the corresponding parameter PDFs derived from the global long-term ~~dataset PDFs can be used to determine an appropriate~~

~~sample. The training sample~~data set. Data from the month of January 2008 is used to determine the optimal number of classes (k) and fuzzy exponent (ϕ) 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 ~~one~~ the same month of global observations (January 2008) ~~as our 'standard full dataset' for the analysis.~~ to do the subsequent comparisons with COCA results.

Figure 2 ~~(panels a, b, & c)~~ shows approximate probability density functions (computed by normalizing the sum of the occurrence frequencies ~~of~~to 1) for different lidar ~~observations~~observables for liquid water clouds, randomly- and horizontal-oriented ice clouds, and aerosols during all of January 2008. Liquid water clouds have the largest ~~backscatter coefficients~~$<\beta'_{532}>$ and ~~color ratios~~χ' values compared with other species. Aerosols generally have the smallest ~~color ratios, depolarization ratios~~χ', δ_v, and ~~backscatter coefficient,~~$<\beta'_{532}>$, and ice clouds have the largest ~~depolarization ratios~~δ_v compared with the other two species. There ~~are overlaps~~is overlap between species, but these three parameters are still sufficient to separate aerosols and different phases of cloud in most cases. The three bottom panels (d, e, & f) in Figure 2 are from a single half orbit (2008-09-06T01-35-29ZN) of observations. The PDFs of one half orbit and one month of observations appear to agree very well, which means that focused feature clustering ~~with~~studies that use the FKM method can be ~~trained using~~applied to a small sample ~~as small~~such as one-half orbit of observations and not cause significant biases to the standard full dataset.

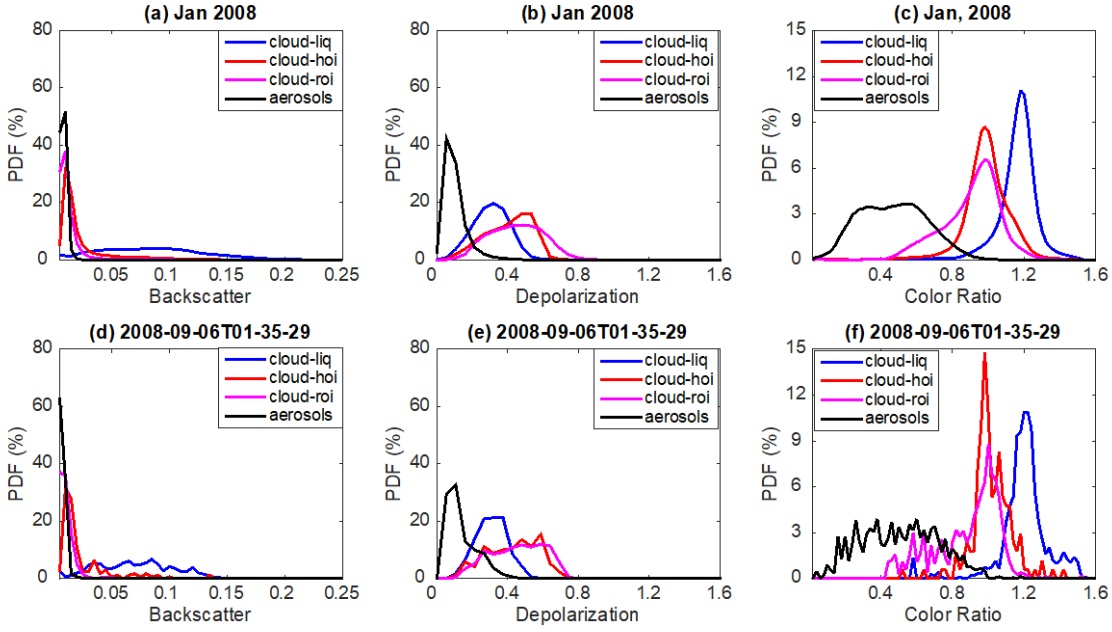


Figure 2: Comparisons of approximate probability density functions computed by normalizing the sum of the occurrence

frequencies to 1; top row (panels a – c) shows data from all of January 2008; bottom row (panels d – f) shows data from a single half-orbit (06 Sep. 2008, 01:35:29 GMT). The left column (panels a and d) compares total attenuated backscatter PDFs; the center column (panels b and e) compares volume depolarization ratio PDFs; and the ~~left~~right column (panels c and f) compares total attenuated backscatter color ratio PDFs. Black lines represent aerosols, blue lines represent liquid water clouds, red lines represent ice clouds dominated by horizontal oriented ice (HOI~~,~~), and magenta ~~line~~lines represent ice clouds dominated by random oriented ice (ROI~~.~~).

~~2.2~~3.3 Data ~~filter~~filtering

We filtered the training data to eliminate outliers in the ~~backscatter, depolarization~~$<\beta'_{532}>$, δ_v, and ~~color ratio~~χ' 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 ~~chosen~~selected filter thresholds ~~are summarized~~retain more than 98% of all features within the original data set. A summary of the thresholds is given in Table 1. The selection of these thresholds is based on the PDFs shown in Figure 2.

Table 1: Filter thresholds for FKM lidar observables

~~Dimension of the satellite observations~~Lidar observable	Filter criteria
~~1.~~Mean ~~Layer Backscatter (AB)~~attenuated backscatter at 532 nm	$0 \leq$ ~~AB~~ $<\beta'_{532}> \leq 0.2$ sr^{-1} km^{-1}
~~2. Layer~~ Integrated ~~Depolarization (DR)~~volume depolarization ratio	$0 \leq$ ~~DR~~$\delta_v \leq 2$
~~3. Layer Integrated Color~~Total attenuated backscatter color ratio ~~(CR)~~	$0 \leq$ ~~CR~~$\chi' \leq 2$

~~Table 1: Thresholds for individual satellite measurements.~~

~~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 ~~have spherical shape~~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 ~~is thus the best for FKM CAD analysis.~~when the clusters are ellipsoidal-shape clouds. The Mahalanobis distance (d_{ij}) of an observation i from a set of observations (x_{il}) with centers c_{jl} (x_{il} - c_{jl} is an l-dimensional vector) is defined in Eq. 8 (Mahalanobis, 1936).

~~$d_{ij}^2 = (x_{il} - c_{jl})^T S^{-1} (x_{il} - c_{jl});\ i = 1, \ldots n;\ j = 1, \ldots k;\ l = 1, \ldots, p$~~ .

$$d_{ij}^2 = (x_{il} - c_{jl})^T S^{-1} (x_{il} - c_{jl}), i = 1,2,\ldots,n; j = 1,2,\ldots,k; l = 1,2,\ldots,p \quad (8)$$

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. ~~When the covariance matrix is the identity matrix, the Mahalanobis distance becomes the Euclidean distance.~~ If 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:

~~2.4~~ $\langle \beta'_{532} \rangle = \frac{\sum_z \beta'_{532,\parallel}(z) + \sum_z \beta'_{532,\perp}(z)}{z}$, , and .

$$\langle \beta'_{532} \rangle = \frac{1}{N} \sum_{n=1}^{N} \beta'_{532,\parallel}(z_n) + \beta'_{532,\perp}(z_n), \quad \delta_v = \sum_{n=1}^{N} \beta'_{532,\perp}(z_n) \Big/ \sum_{n=1}^{N} \beta'_{532,\parallel}(z_n) \quad \text{, and} \quad \chi' = \langle \beta'_{1064} \rangle \Big/ \langle \beta'_{532} \rangle = \sum_{n=1}^{N} \beta'_{1064}(z_n) \Big/ \sum_{n=1}^{N} \beta'_{532}(z_n) .$$

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 ϕ

The selection of an optimal number of classes k ($1 < k < n$) and degree of fuzziness ϕ ($\phi > 1$) 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 ~~and~~to validate the results ~~validated~~. Among different validation functions, the fuzzy performance index (FPI) and the modified partition entropy (MPE) are considered two of the most useful ~~indexes~~indices among seven examined by Roubens (1982) to evaluate the effects of varying class number. The FPI is defined as in Eq. 9, where F is the partition coefficient calculated from Eq. 10. The MPE is defined as in Eq. 11, with the entropy function (H) calculated from Eq. 12.

The ideal number of continuous and structured classes (k) can be established by simultaneously minimizing both ~~these two~~

~~measures (~~FPI and MPE~~).~~. For the fuzziness exponent, if the value of ϕ is too low~~,~~ the classes ~~are~~become more discrete and the membership values either approach 0 or 1. But if ~~the value~~ 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. 13, Bezdek, 1981) decreases with increasing of both fuzzy exponent (ϕ) and the number of classes (k). They plotted a series of objective functions versus the fuzzy exponent (ϕ) for a given class where the best value of ϕ for that class is at the first maximum of objective function curves (Odeh et al. 1992a, McBratney and Moore 1985). Therefore, choosing an optimal combination of class number (k) and fuzzy exponent (ϕ) is established on the basis of minimizing both values of FPI and MPE and the least maximum of the objective function.

~~$FPI = 1 - \frac{k \times F - 1}{k - 1}$~~

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

and (9)

~~$F = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^2$~~ (10)

~~MPE = $\frac{H}{log k}$~~

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

(10)

$$MPE = \frac{H}{\log k}$$, and (11)

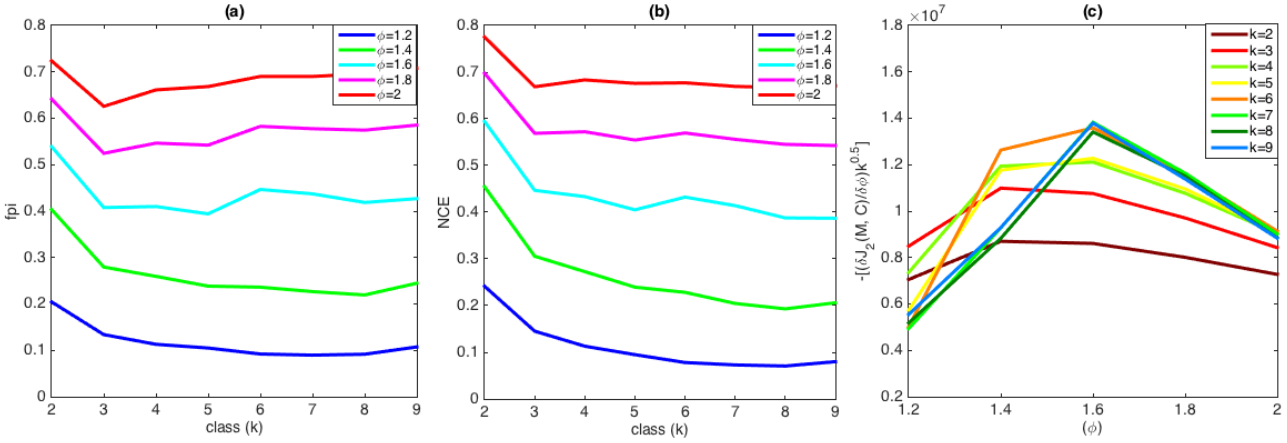
~~$H = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij} \times \log(m_{ij})$~~ (12)

~~$\frac{\delta J(M,C)}{\delta \phi} = \sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}^{\phi} \log(m_{ij}) d_{ij}^{2}$ (13)~~

$$H = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij} \times \log(m_{ij}) \quad ; \qquad (12)$$

$$\frac{\delta J(M,C)}{\delta \phi} = \sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}^{\phi} \log(m_{ij}) d_{ij}^{2} \quad . \qquad (13)$$

Using one month of ~~"no clear scenes" data~~layer optical properties reported in the CALIOP level 2 merged layer products, we created Figure 3 to determine optimal values for k and ϕ. From this figure, we ~~concluded~~conclude that the ideal number classes for CALIOP layer classification is either 3 or 4, with corresponding fuzzy exponents equal to 1.4 or 1.6 (we use 1.4 for the analyses 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 3 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 ~~3~~2 or ~~2~~3 and ϕ equal to 1.4.



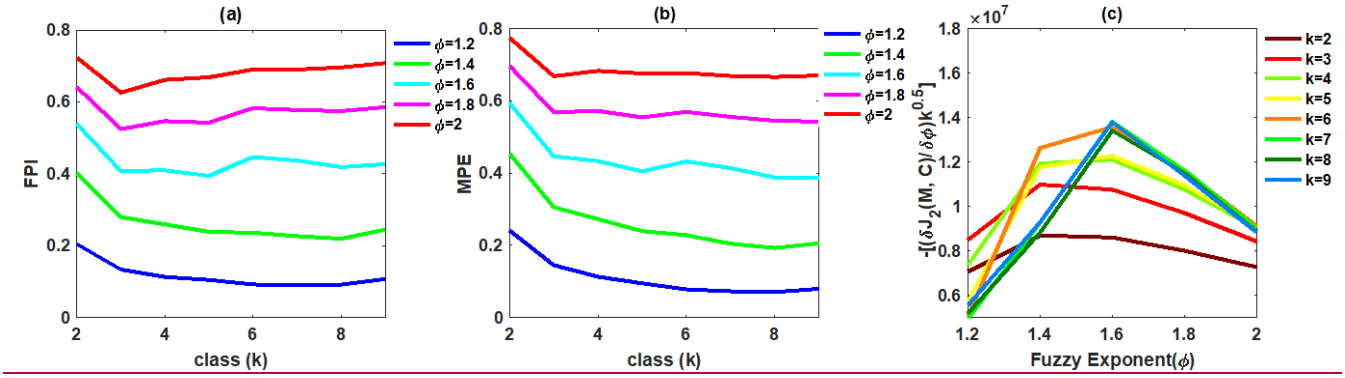


Figure 3: Determination of the number of classes ~~number~~, k, and the fuzzy exponent, ϕ, for the FKM cloud ~~and~~ aerosol discrimination algorithm: (a) FPI (y-axis) versus class number k (x-axis) for different values of fuzzy exponent ϕ; (different colors); (b) MPE (y-axis) versus class number k (x-axis) for different values of fuzzy exponent ϕ; (different colors); and (c) objective function values (y-axis) versus the fuzzy exponent ϕ (x-axis) for various class numbers. (different colors).

~~3.~~4 Cluster results and comparison with V3 and V4 data

~~34~~.1 CAD from the Fuzzy K-Means algorithm

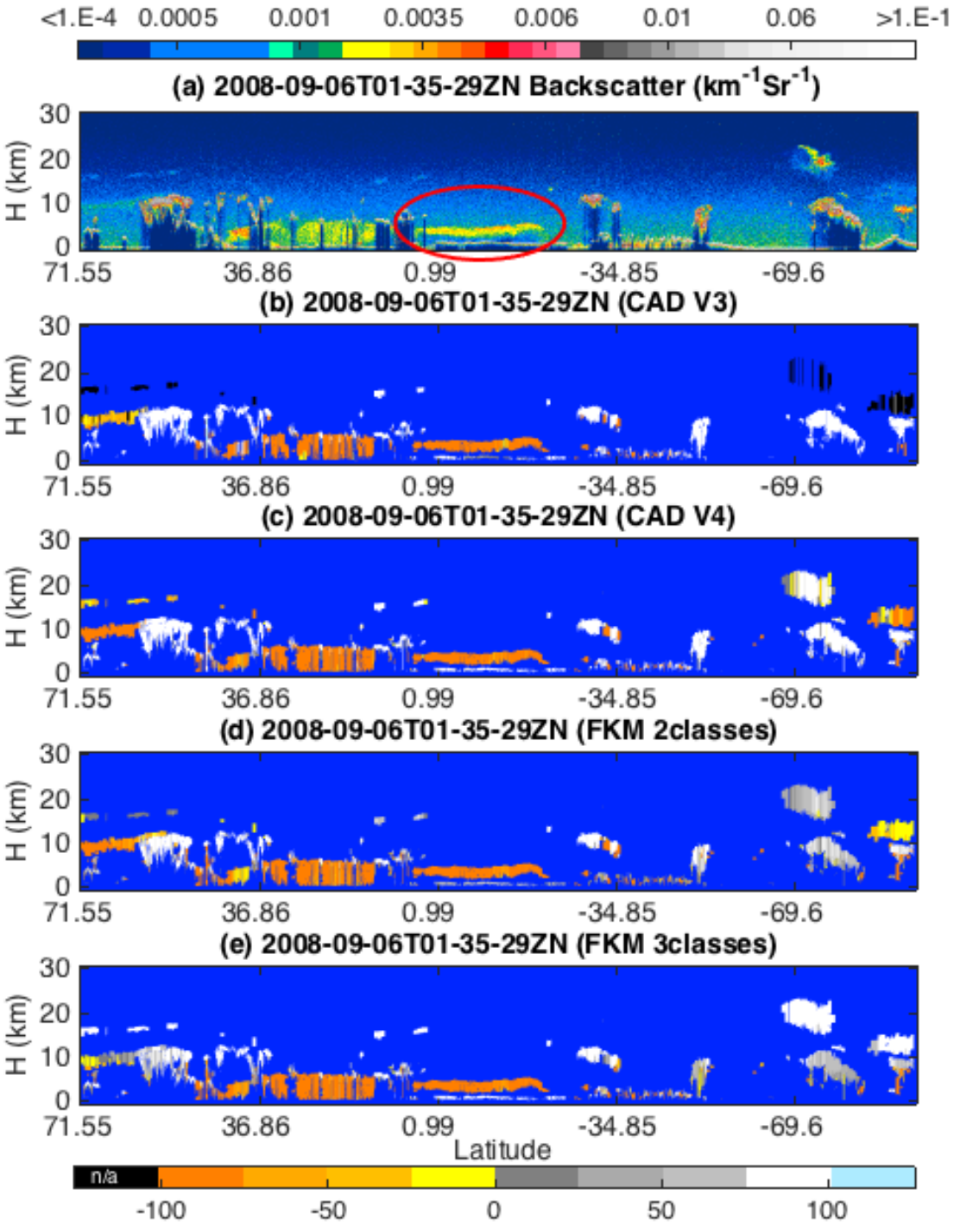
According to Liu et al. (2009), the CAD ~~value~~score for any layer is the difference between the probability of being a cloud and the probability of being an aerosol (Eq. 14). We calculate the ~~fuzzy K-means~~FKM CAD score in a similar way, where the ~~probability is derived from the~~COCA probabilities are replaced with FKM membership values. For the 3-class FKM analyses, the cloud membership value is the sum of memberships of ice and water clouds (two classes). The ~~fuzzy K-means~~FKM CAD ~~of clouds and aerosols~~score is found using

~~$$CAD_{FKM} = \frac{M_{cloud} - M_{aerosol}}{M_{cloud} + M_{aerosol}} \times 100. \quad (14)$$~~

$$CAD_{FKM} = \frac{M_{cloud} - M_{aerosol}}{M_{cloud} + M_{aerosol}} \times 100 \quad . \quad (14)$$

Figure 4 compares the operational V3 and V4 CAD products and our CAD_{FKM} classifications for a single nighttime orbit segment (06 September 2008, beginning at 01:35:29 GMT). Generally speaking, CAD_{FKM} from both the 2-class and 3-class analyses are quite similar to both the V3 and V4 ~~operational CAD~~ COCA values. When COCA CAD ~~values~~scores are positive (namely clouds, shown in whitish colors in Fig. 4) in V3 and V4, the 2-class and 3-class CAD_{FKM} values are also positive.

Likewise, when COCA CAD ~~values~~scores are negative (namely aerosols, yellowish colors in Fig. ~~4) in the operational data,~~4), the 2-class and 3-class CAD_{FKM} values are also negative. Furthermore, the particular orbit selected here includes ~~some~~the observations ~~(latitudes between 0° and 20°S)~~ of a plume of high, dense smoke lofted over low water clouds~~.~~ (latitudes between 0° and 20°S within the red oval). For these water clouds beneath dense smoke, both the V3 operational CAD and the 2-class CAD_{FKM} label them as clouds with low positive values. On the other hand, the V4 operational CAD and the 3-class CAD_{FKM} return higher values much closer to 100. The reasons for these differences will be discussed in ~~section 3~~Sect. 5.2 and ~~3~~Sect. 5.4. Note too that weakly scattering edges of cirrus clouds (hereafter, cirrus fringes) ~~around 74~~beyond 69.6°S are misclassified as aerosols by both the 2-class and 3-class CAD_{FKM} (Figure 4c and d) but are correctly classified as cloud by the operational V4 algorithms.


<1.E-4 0.0005 0.001 0.0035 0.006 0.01 0.06 >1.E-1
(a) 2008-09-06T01-35-29ZN Backscatter (km-1Sr-1)
(b) 2008-09-06T01-35-29ZN (CAD V3)
(c) 2008-09-06T01-35-29ZN (CAD V4)
(d) 2008-09-06T01-35-29ZN (FKM 2classes)
(e) 2008-09-06T01-35-29ZN (FKM 3classes)
H (km)
30
20
10
0
71.55
36.86
0.99
-34.85
-69.6
Latitude
n/a
-100
-50
0
50
100

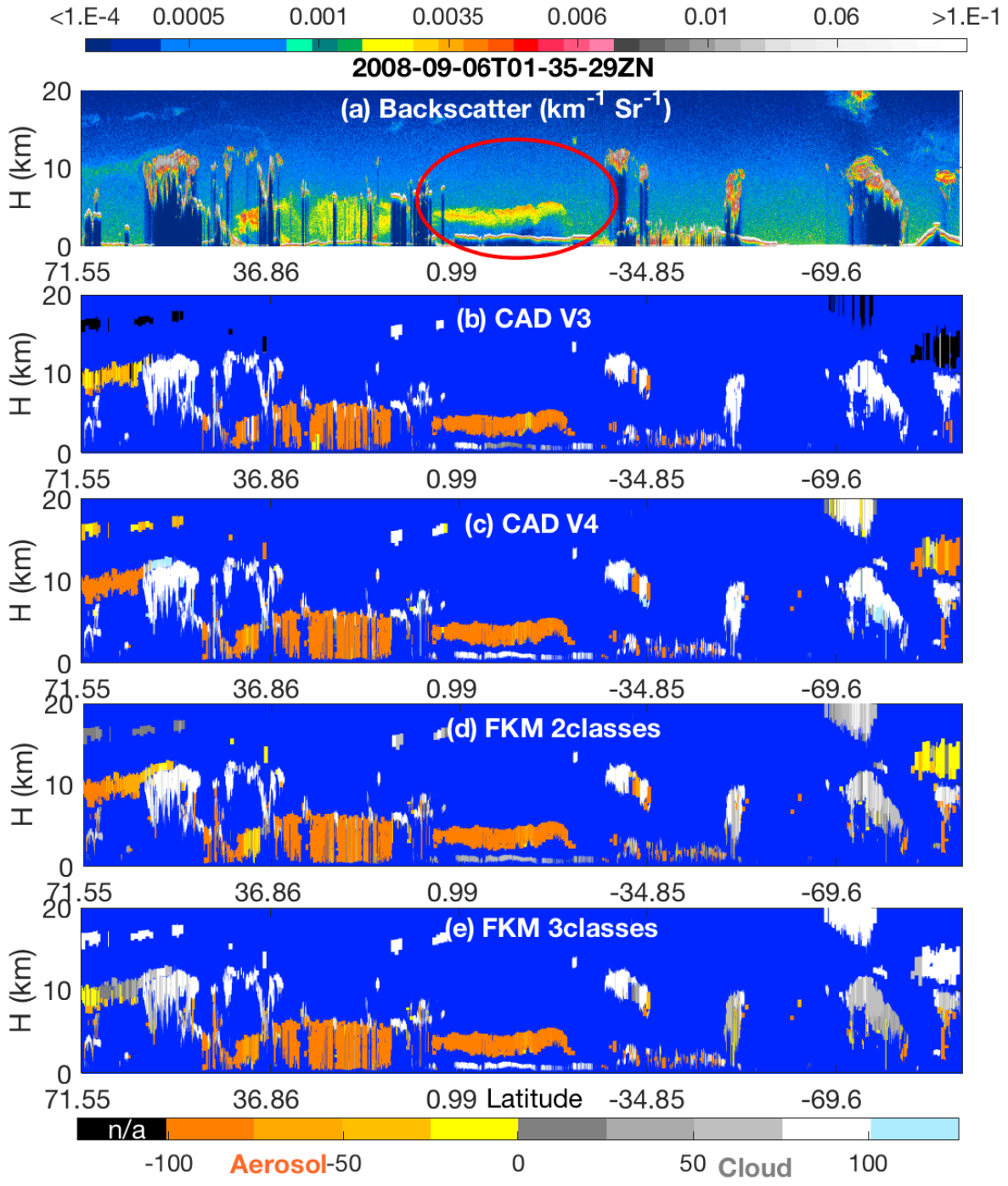

<1.E-4
0.0005
0.001
0.0035
0.006
0.01
0.06
>1.E-1
2008-09-06T01-35-29ZN
(a) Backscatter (km-1 Sr-1)
(b) CAD V3
(c) CAD V4
(d) FKM 2classes
(e) FKM 3classes
H (km)
20
10
0
71.55
36.86
0.99
-34.85
-69.6
Latitude
n/a
-100
Aerosol
-50
0
50
Cloud
100

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 2-class FKM CAD algorithm, and (e) the 3-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 (panels b–e), stratospheric layers are shown in black and, cirrus fringes are shown in pale blue, and regions of "clear air" where no features were detected are shown in pure blue. Latitude units in degrees; positive: north, negative: south.

34.2 Uncertainties: class overlap

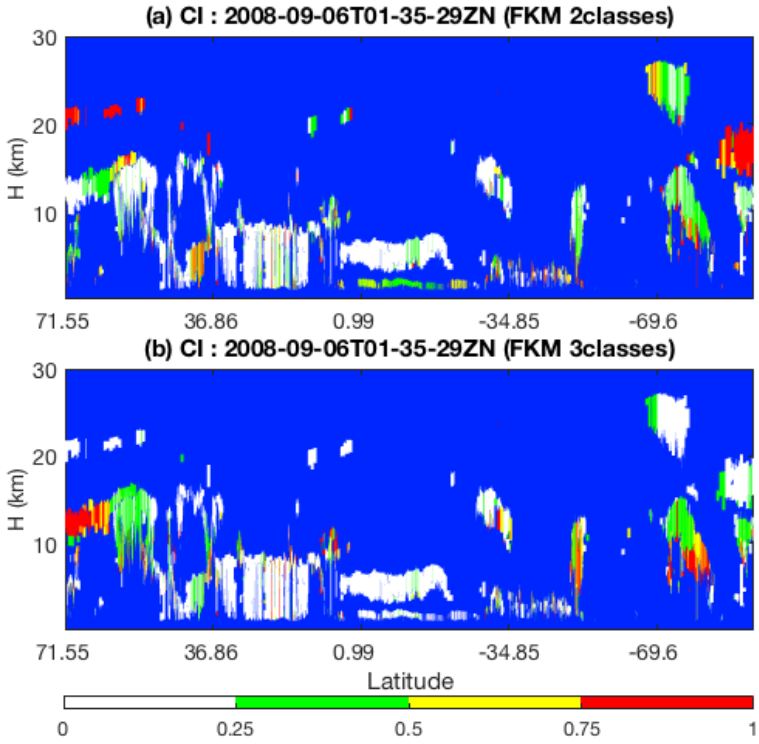
The confusion index (CI) is a measure of the degree of class overlap or cluster uncertainty between classes (Burrough and McDonnell, 1998). In effect, it measures how wellconfidently each individual observation has been classified. CI values are calculated from Eq. 15, where m_{max} denotes the biggest membership value and m_{max-1} is the second biggest membership value for each individual observation (i):

$$CI = [1 - (m_{max_i} - m_{(max-1)_i})] \quad (15)$$

CI value approaches zero when m_{max} is much larger than m_{max-1}, indicating that the observation is more likely to belong to one dominant class. CI approaches one when m_{max} is almost equal to m_{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- CI) × 100 for the 2-class FKM algorithm is equivalent to the absolute value of the CAD_{FKM} score.

Figure 5 shows CI values for 2-class and 3-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 2-class and 3-class CAD_{FKM} classifications. The exceptions are stratospheric features (mostly near polar regions), cloud fringes, high altitude aerosols and, for 2-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 2-class CAD_{FKM} are larger than 0.5. However, the 3-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 3 classes are used. For cloud fringes, the CI values are high for both the 2-class and 3-class CAD_{FKM}. According to the CAD_{FKM} results, cirrus fringes are essentiallysomewhat different from the neighboring portions of the cirrus and more closely resemblelayer, as they also bear some similarity to the dust particles that are the predominant sources of ice nuclei (DeMott et

al., 2010).


(a) CI : 2008-09-06T01-35-29ZN (FKM 2classes)
H (km)
30
20
10
71.55
36.86
0.99
-34.85
-69.6
(b) CI : 2008-09-06T01-35-29ZN (FKM 3classes)
H (km)
30
20
10
71.55
36.86
0.99
-34.85
-69.6
Latitude
0
0.25
0.5
0.75
1

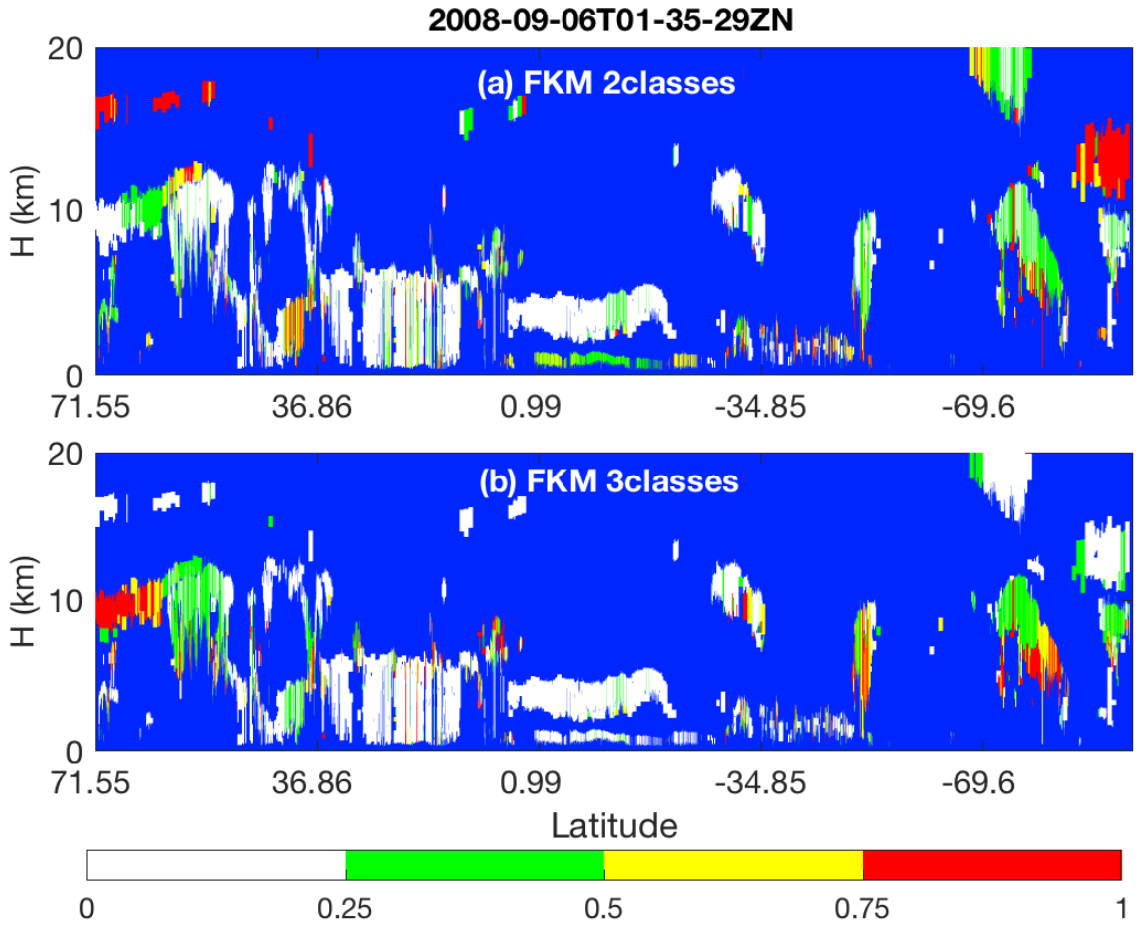
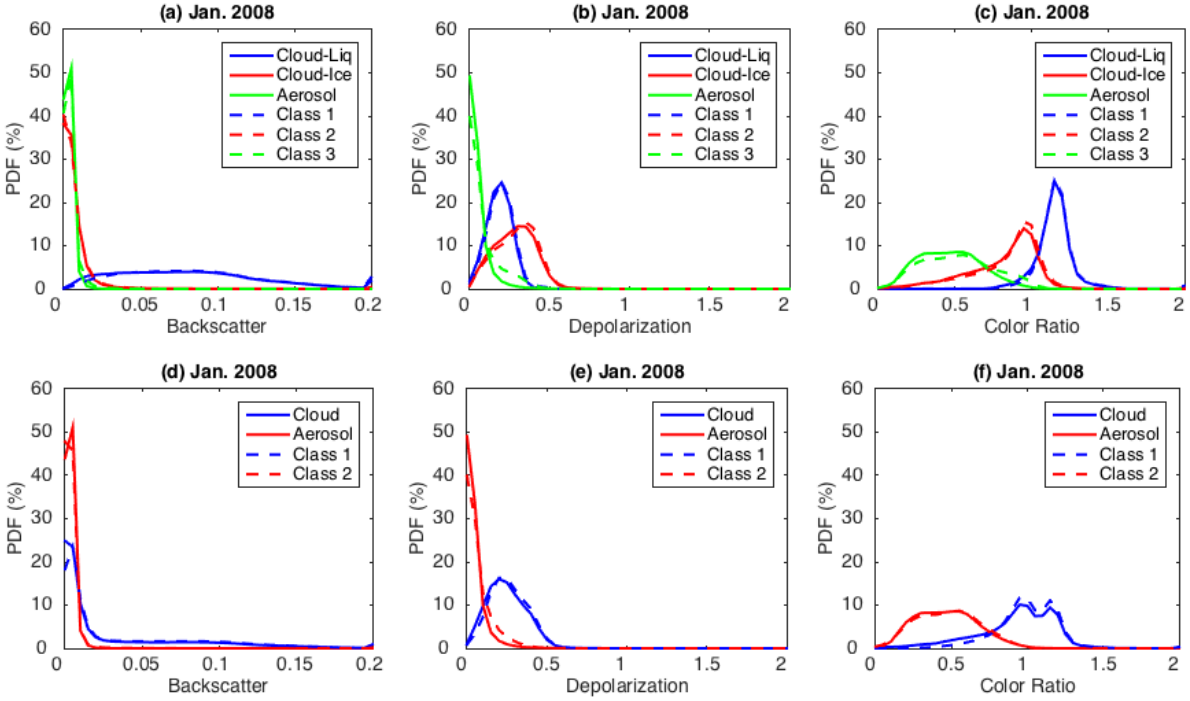


Figure 5: for the same data shown in Figure 4, the upper panel (a) shows the confusion index for 2-class CAD_{FKM}, and the lower panel (b) shows the confusion index for 3-class CAD_{FKM}. The pure blue color once again indicates those regions where no atmospheric layers were detected.

4.3~~.3 Statistic~~ Statistical comparisons of clouds and aerosols

In this sub-section, we present statistical analyses of our results for all of January 2008, followed by explorations of individual case studies in the next sub-section. We first compare the PDFs of the different lidar optical parameters used in the 2-class and 3-class CAD_{FKM} classifications to the PDFs of those same parameters derived for the ~~operational CAD~~COCA classifications (~~Figure~~Fig. 6). ~~It~~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 ~~backscatter coefficient, depolarization ratio~~$<\beta'_{532}>$, δ_v, and ~~color ratio of~~χ' that characterize the clouds and aerosols ~~from~~determined by the FKM classifications agree well with the PDFs from the V4 CAD classifications. ~~-~~Figures 6d, 6e, and 6f compare the 2-class CAD_{FKM} results ~~(dashed lines)~~ to the operational algorithm ~~(solid lines).~~ In these figures, the PDFs of ~~backscatter~~ $<\beta'_{532}>$ (Fig. 6d), ~~depolarization~~δ_v (Fig. 6e) and ~~color ratio~~χ' (Fig. 6f) of FKM class 1 (blue dashed lines) agree well with those of V4 cloud PDFs (blue solid lines), while the PDFs of these different parameters of FKM class 2 (red dashed lines) agree well with those of V4 aerosol (red solid lines) PDFs. Figures 6a, 6b, and 6c compare the 3-class CAD_{FKM} results to the operational algorithm. Once again, the comparisons are quite good: the ~~PDFs~~shapes of the PDFs of FKM class 1 (blue dashed) agree well with the V4 water cloud (blue solid) PDFs, while the PDFs of FKM class 2 (red dashed) and 3 (green dashed) individually agree well with, respectively, the V4 ice cloud (red solid) and aerosol (green solid) PDFs. The class means for $<\beta'_{532}>$ are smallest for aerosols/class 3 (0.0034 ± 0.0022 ($km^{-1}sr^{-1}$) and 0.0041 ± 0.0193 ($km^{-1}sr^{-1}$), respectively) and slightly larger for ice clouds/ class 2 (0.0075 ± 0.0086 ($km^{-1}sr^{-1}$) and 0.0062 ± 0.0183 ($km^{-1}sr^{-1}$), respectively). Water clouds/class 1 have the largest $<\beta'_{532}>$ mean values (0.0804 ± 0.0526 ($km^{-1}sr^{-1}$) and 0.0850 ± 0.0454 ($km^{-1}sr^{-1}$), respectively). For δ_v, the largest mean values are found for ice clouds/ class 2, followed by water clouds/class 1 and then aerosol/class 3. Class mean χ' 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.



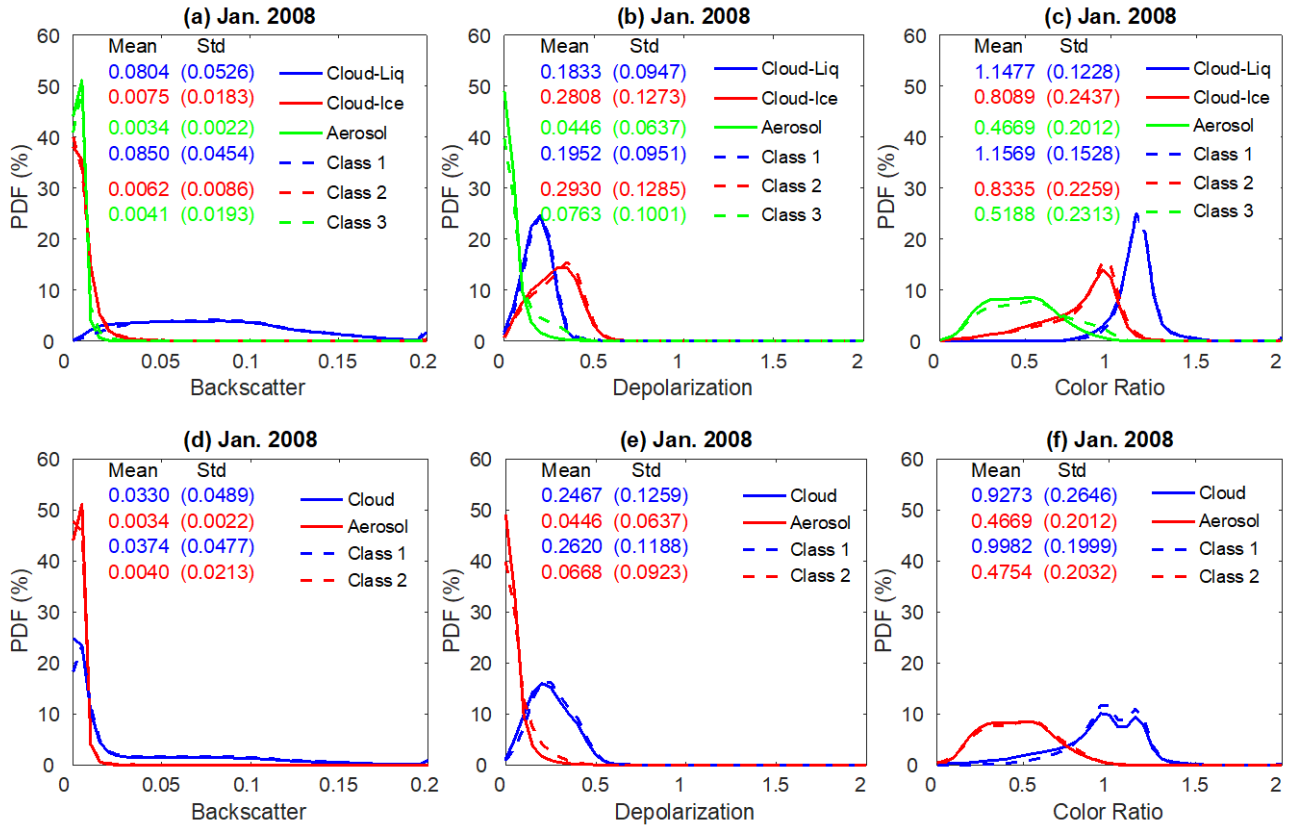


Figure 6: ~~probability density functions~~PDFs derived from all data from January 2008. The top row compares V4 operational CAD PDFs to the PDFs derived from CAD_{FKM} 3-class results. V4 CAD PDFs for liquid water clouds, ice clouds, and aerosols are plotted in, respectively, solid blue, red and green lines. Similarly, CAD_{FKM} 3-class PDFs for classes 1, 2, and ~~three~~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_{FKM} 2-class results, where once again the V4 CAD PDFs are shown in solid lines and the CAD_{FKM} 2-class PDFs are shown in dashed lines. ~~Attenuated backscatter~~ PDFs of $<\beta'_{532}>$ are shown in the left column (panels a and d), ~~depolarization ratio~~δ_v PDFs in the center column (panels b and e), and ~~color ratio~~χ' PDFs in the right column (panels c and f).

Figure 7 compares ~~one month (January 2008) of~~ the geographical (~~Fig. 7,~~ panels a-f) and zonally-averaged (~~Fig. 7,~~ panels g-1) distributions of 2-class CAD_{FKM} occurrence frequencies to the COCA cloud and aerosol occurrence frequencies ~~derived from the operational V4 CAD classifications, and the differences between V4 and FKM~~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 inter-tropical 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 (upper two rows of ~~Figure~~Fig. 7). In the zonal mean plots (lower two rows of ~~Figure~~Fig. 7), 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 ~~CAD~~ cloud distributions. Likewise, the distributions of FKM class 2 closely resemble the COCA V4 ~~CAD~~ aerosol distributions. ~~Large~~Looking at the difference plots (righthand column of Fig. 7), some fairly large differences ~~appear~~are seen in ~~Polar Regions~~the polar regions, where the composition and intermingling of clouds and aerosols is notably different from other regions of the globe ~~(last column of Figure 7). In summary,~~. 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 ~~show~~presented here establish that, excluding the polar regions, the cloud-aerosol discrimination derived using an unsupervised FKM method is ~~largely~~statistically consistent with the classifications produced by the operational V4 CAD algorithm ~~beyond Polar Regions~~.

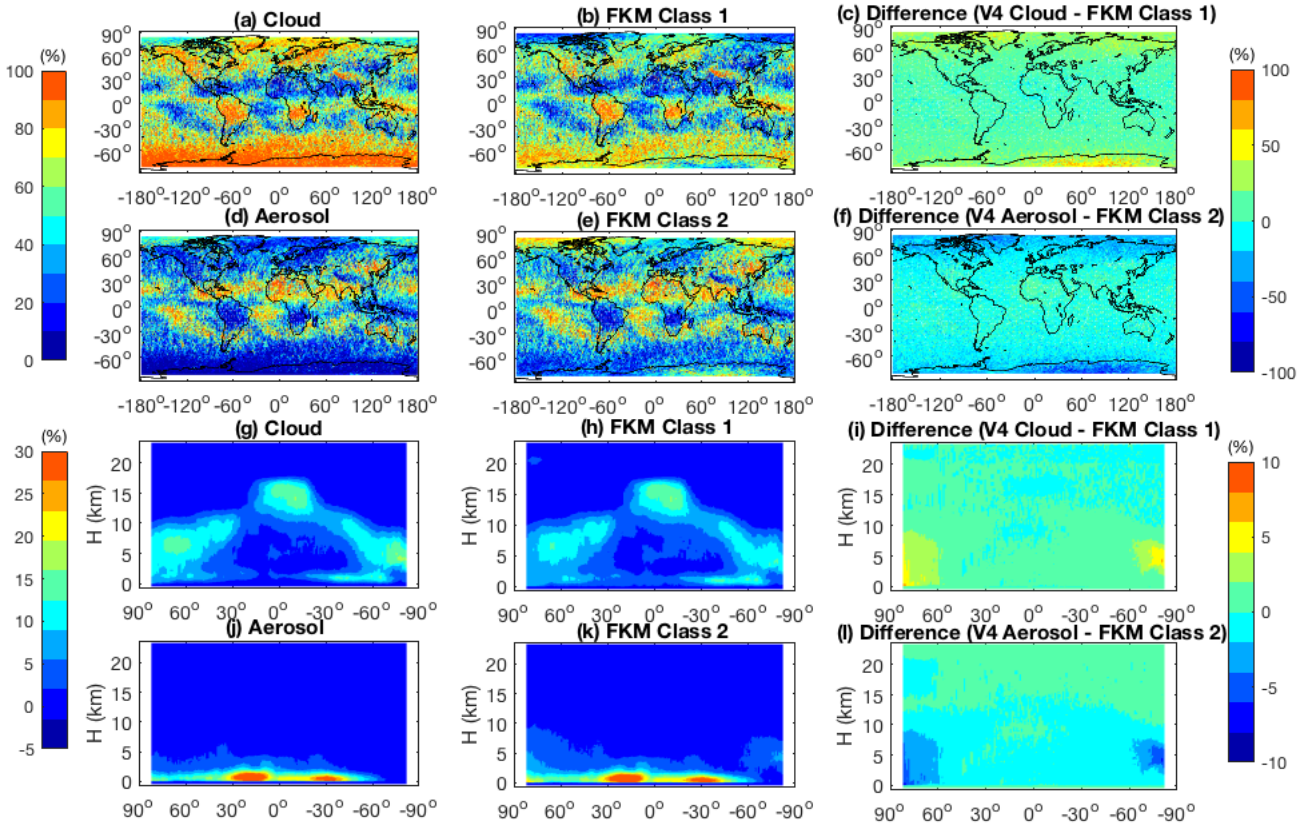


Figure 7: distributions of feature type occurrence frequencies during January 2008. Panels in the left column show V4 ~~operational CAD~~COCA results; panels in the center column show CAD_{FKM} 2-class results; and the panels in the right column show the ~~absolute~~percentages of differences between the left and center columns. The top two rows show maps of occurrence frequencies as a function of latitude and longitude for clouds (panels a–c) and aerosols (panels d–f). The bottom two rows show the zonal mean occurrence frequencies of clouds (panels g–i) and aerosols (panels j–l).

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 ~~used~~construct confusion matrices, which use the January 2008 5-km ~~Layer~~merged layer data between 60°S and 60°N to calculate the concurrent frequency of cloud and aerosol identifications made by the ~~operational V4 5km CAD algorithm~~COCA and CAD_{FKM} algorithms. We summarize the occurrence frequency statistics in Table 2. From the table we find that for our test month COCA V3 ~~CAD~~ agrees with COCA V4 CAD for 96.6 % of the cases. The agreements are around 90% for the entire globe including regions beyond 60° (not shown here). The FKM 2-class and 3-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 2-class CAD_{FKM}, results from the 3-class CAD_{FKM} agree somewhat better with the classifications from both the V3 and V4 CAD algorithms. Consistent with previous results in this paper, the 3-class CAD_{FKM} appears better able to separate clouds and aerosols than the 2-class CAD_{FKM}. Figure 4 provides an additional example. For those water clouds beneath dense smoke, the 3-class CAD_{FKM} scores are substantially higher than both the 2-class CAD_{FKM} scores and the operation V3 CAD scores, indicating that the 3-class CAD_{FKM} algorithm correctly identifies these features with much higher classification confidence. 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.

~~Agreement (%)~~		~~V3~~			~~V4~~			~~FKM (2-classes)~~			~~FKM (3-classes)~~		
		~~C~~	~~A~~	~~T~~	~~C~~	~~A~~	~~T~~	~~C~~	~~A~~	~~T~~	~~C~~	~~A~~	~~T~~
~~V3~~	~~C~~	-			~~66.1~~	~~2.1~~		~~63.2~~	~~5.4~~		~~65.1~~	~~3.2~~	
	~~A~~				~~1.2~~	~~30.5~~		~~1.0~~	~~30.4~~		~~1.9~~	~~29.8~~	
	~~T~~						~~96.6~~			~~93.6~~			~~94.9~~
~~V4~~	~~C~~	-			-			~~58.9~~	~~5.3~~		~~60.9~~	~~3.2~~	
	~~A~~							~~1.5~~	~~34.4~~		~~2.5~~	~~33.4~~	
	~~T~~									~~93.2~~			~~94.3~~
~~FKM~~	~~C~~										~~63.2~~	~~5.1~~	

(2-classes)	A	-	-	-	1.4	30.4	
	T						93.6

Table 2:2: Statistical confusion matrix of a 1-month (Jan. 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).

3

Agreement (%)		V4			FKM (2-classes)			FKM (3-classes)		
		C	A	T	C	A	T	C	A	T
V3	C	66.1	2.1		63.8	4.5		64.6	3.7	
	A	1.2	30.5		1.1	30.6		1.9	29.9	
	T			96.6			94.4			94.5
V4	C	-			59.3	4.9		60.6	3.6	
	A				1.5	34.4		2.4	33.4	
	T						93.6			94.0
FKM (2-classes)	C	-			-			60.1	0.6	
	A							2.8	36.5	
	T									96.7

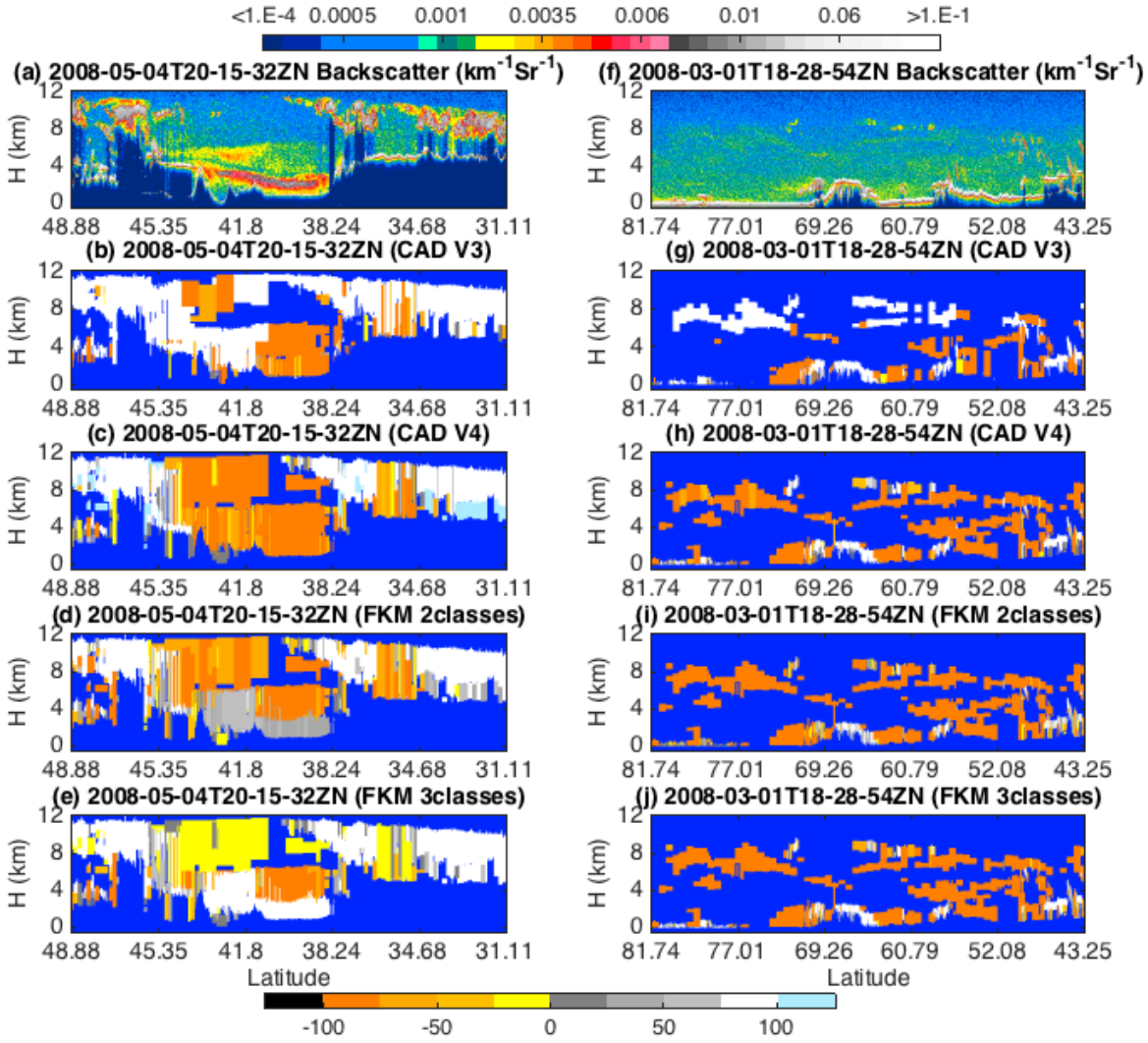
4.4 Special cases study

~~We see from section 3.3 that, statistically, the CAD$_{FKM}$ classifications agree well with the operational V3 and V4 CAD results.~~ In this section~~,~~ we investigate several of the challenging classification cases ~~for which~~that motivated the extensive changes made in COCA in the ~~classifications disagree between~~ transition from V3 ~~and~~to V4 ~~CADs~~ (Liu et al. 2018). 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 Figure 4, the CAD$_{FKM}$ algorithm, like the operational CAD algorithm, ~~also has~~can occasionally have difficulty correctly identifying high altitude smoke, dense dust, lofted dust, cirrus fringes, polar stratosphere clouds (PSC) and stratospheric volcanic ash (Figure 8-10). We briefly review each of these cases below.

a. Dust

Two different dust cases are selected for this study (~~Figure~~Fig. 8). ~~One dust~~The first case ~~located~~examines nighttime measurements of a deep and sometimes extremely dense dust plume in the Taklamakan desert~~, is a night time granule~~ beginning at 20:15:32 UTC on 4 May 2008, as shown in ~~Figures~~Figs. 8 a-e. The second case ~~is lofted~~ investigates spatially diffuse Asian dust ~~from~~lofted high into the atmosphere while being transported toward the Arctic during a ~~night time granule~~

~~at~~nighttime orbit segment beginning at 18:28:54 UCT on ~~the~~ 1 March 2008, as shown in ~~Figure~~Figs. 8 f-j. ~~These dust~~ 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 gray scale, with brighter and whiter hues indicating higher classification confidence. Dust layers in Taklamakan exhibit ~~strong~~high 532 nm attenuated backscatter ~~coefficient~~coefficients, high depolarization ~~ratio and color~~ ratios ~~and~~(not shown), and attenuated backscatter color ratios close to 1 (also not shown). As seen between ~44° N and ~40° N, layers with this combination of layer optical properties are frequently misclassified as ice clouds in ~~V3 but~~ COCA V3 (Fig. 8b). However, in COCA V4 these same layers are much more likely to be correctly classified as aerosol ~~in V4. The~~ (Fig. 8c).The 2-class and 3-class CAD_{FKM} classifications ~~for the 2-class and 3-class~~ both agree with ~~version 3 but with~~ COCA V4 for the lofted aerosols, but misclassify the densest portions of the dust plume as low confidence ~~in the dense dust case, suggesting that FKM does not~~ cloud. For the lofted Asian dust case shown in Figure 8 f-j, COCA V3 frequently misclassifies dust filaments as cloud, whereas COCA V4 correctly ~~identify the dense dust. But both FKM~~identifies the vast majority as dust. (Note too that many more layers are detected in V4 as a consequence of the changes made to the CALIOP 532 nm calibration algorithms ~~agree well with the version 4 CAD in the lofted Asian dust case.~~(Kar et al., 2018; Getzewich et al., 2018; Liu et al., 2018).) The 2-class and 3-class CAD_{FKM} classifications are essentially identical to those determined by COCA V4, but show higher confidence values for the aerosol layers.


<1.E-4 0.0005 0.001 0.0035 0.006 0.01 0.06 >1.E-1
(a) 2008-05-04T20-15-32ZN Backscatter (km-1Sr-1)
(f) 2008-03-01T18-28-54ZN Backscatter (km-1Sr-1)
(b) 2008-05-04T20-15-32ZN (CAD V3)
(g) 2008-03-01T18-28-54ZN (CAD V3)
(c) 2008-05-04T20-15-32ZN (CAD V4)
(h) 2008-03-01T18-28-54ZN (CAD V4)
(d) 2008-05-04T20-15-32ZN (FKM 2classes)
(i) 2008-03-01T18-28-54ZN (FKM 2classes)
(e) 2008-05-04T20-15-32ZN (FKM 3classes)
(j) 2008-03-01T18-28-54ZN (FKM 3classes)
H (km)
0
4
8
12
48.88 45.35 41.8 38.24 34.68 31.11
81.74 77.01 69.26 60.79 52.08 43.25
Latitude
-100 -50 0 50 100

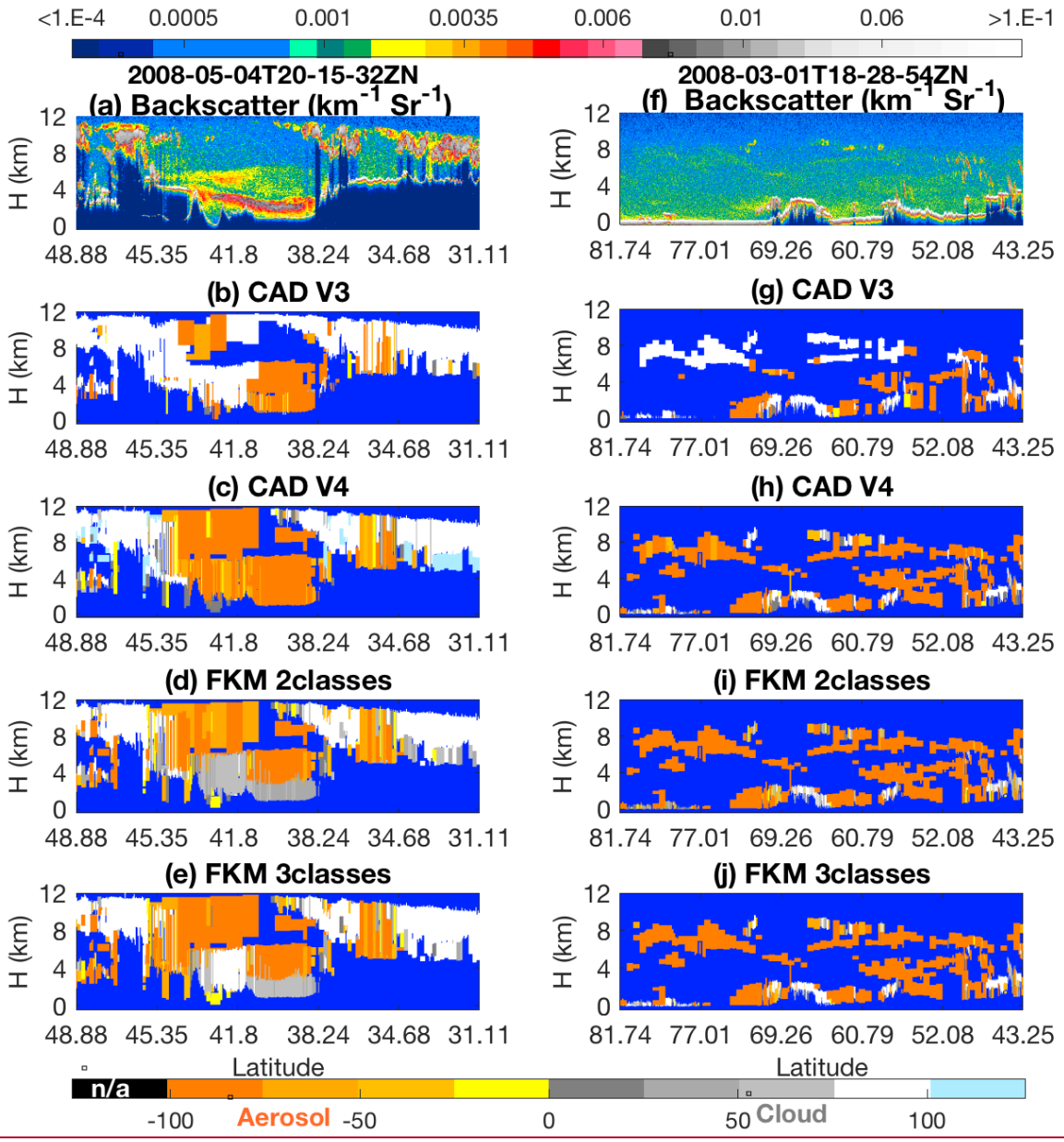
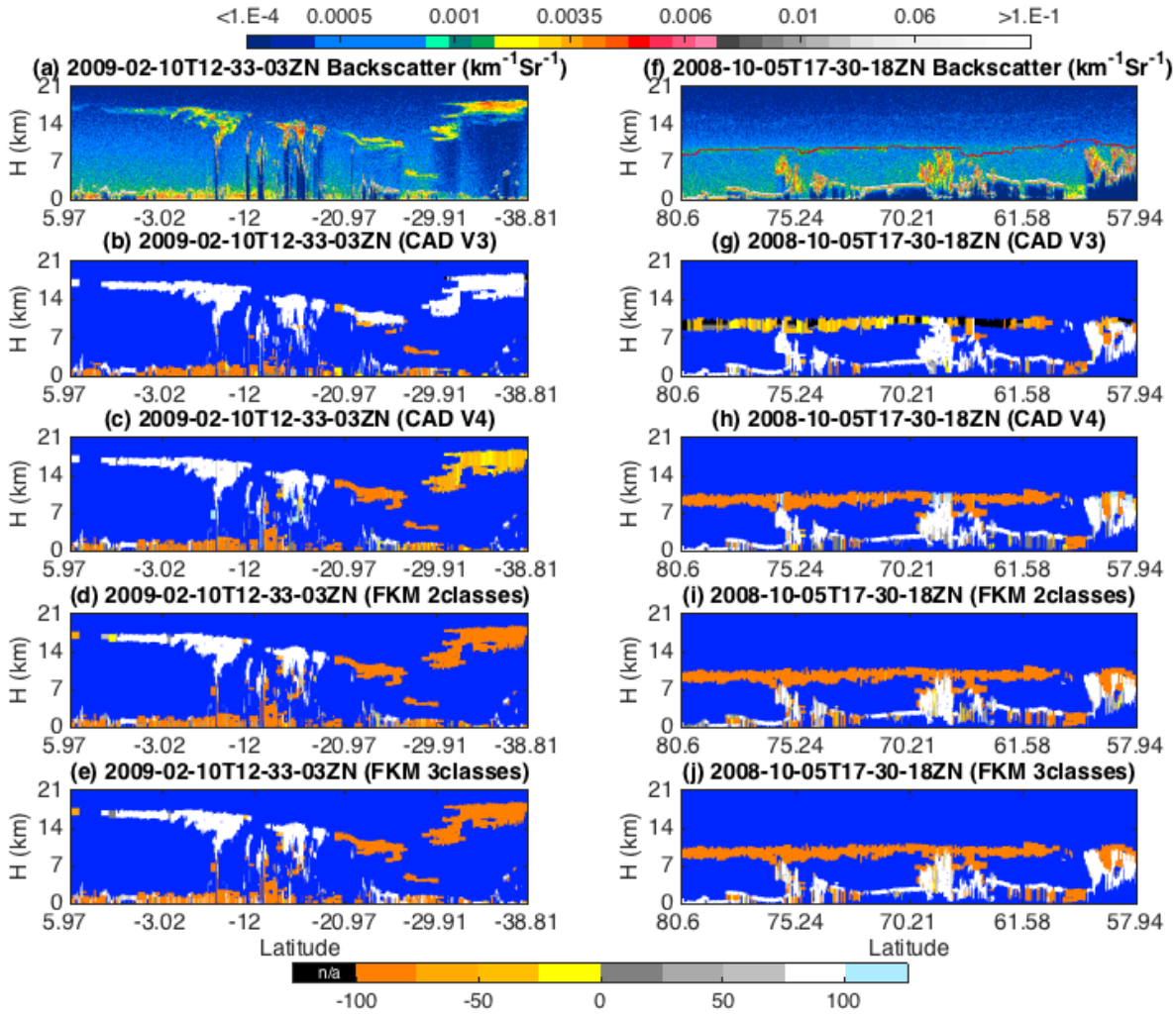


Figure 8: top row shows 532 nm attenuated backscatter coefficients for (a) dust in the Taklamakan basin on 4 May 2008 and (f) lofted Asian dust being transported into the Arctic on 1 March 2008. The rows below show the CAD results reported by four different algorithms; ~~the~~COCA V3 ~~operational CAD~~ (panels b and g), ~~the~~COCA V4 ~~operational CAD~~ (panels c and h), the 2-class CAD_{FKM} (panels d and i), and the 3-class CAD_{FKM} (panels e and j).

b. High altitude smoke

An unprecedented example of high ~~-~~altitude smoke ~~plume~~plumes was observed by CALIPSO during the "Black Saturday" fires that started 7 February 2009 ~~and~~, quickly spread across the Australian state of Victoria~~.~~, and eventually lofted well into the stratosphere (de Laat et al., 2012). Figure 9 ~~(panels f-j)~~ shows extensive smoke layers at 10 km and higher on Monday 10 February between 20°S-40°S. ~~The~~In the V3 ~~CAD algorithm~~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., 2018). Consequently, COCA V3 misclassifies these smoke layers as clouds ~~(or~~ when their base altitude is below the tropopause and as stratospheric features~~), and so too does the FKM 3-class analysis.~~ when the base altitude is higher (Fig. 9b). On the other hand, the V4 CAD ~~and FKM 2-class analysis~~ correctly ~~identify~~identifies them as aerosols (~~not shown here). Without altitude as inputs~~Fig. 9c). In analyzing this scene we used two separate versions of the FKM algorithm. Our standard configuration used z_{mid} as one of the classification attributions, while a second, trial configuration omitted z_{mid}. For the 2-classes FKM, both ~~FKM 2-class and 3-class algorithms identify the layers as aerosols (Figure 9 d, e, also see Figure 11).~~configurations successfully identified the high-altitude smoke as aerosol (Figs. 9d and 9e). But for the 3-classes FKM, including z_{mid} as a classification attribute introduced uniform misclassification of the lofted smoke as cloud (Fig. 9f). However, when z_{mid} is omitted the 3-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~~. The reasons for this~~ , both of which are ~~discussed~~located at similar altitudes and have similar optical properties. Altitude is not a driving factor for classifications, and adds confusion in ~~some detail~~the memberships defined by the Mahalanobis distance (see Eqs. (7) and (8)) in these particular cases. More details are given in section ~~4~~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_{mid} as an input will resolve large fractions of the potential misclassifications from the FKM method.


<1.E-4 0.0005 0.001 0.0035 0.006 0.01 0.06 >1.E-1
(a) 2009-02-10T12-33-03ZN Backscatter (km-1Sr-1)
(f) 2008-10-05T17-30-18ZN Backscatter (km-1Sr-1)
(b) 2009-02-10T12-33-03ZN (CAD V3)
(g) 2008-10-05T17-30-18ZN (CAD V3)
(c) 2009-02-10T12-33-03ZN (CAD V4)
(h) 2008-10-05T17-30-18ZN (CAD V4)
(d) 2009-02-10T12-33-03ZN (FKM 2classes)
(i) 2008-10-05T17-30-18ZN (FKM 2classes)
(e) 2009-02-10T12-33-03ZN (FKM 3classes)
(j) 2008-10-05T17-30-18ZN (FKM 3classes)
H (km)
21
14
7
0
5.97 -3.02 -12 -20.97 -29.91 -38.81
80.6 75.24 70.21 61.58 57.94
Latitude
n/a
-100 -50 0 50 100

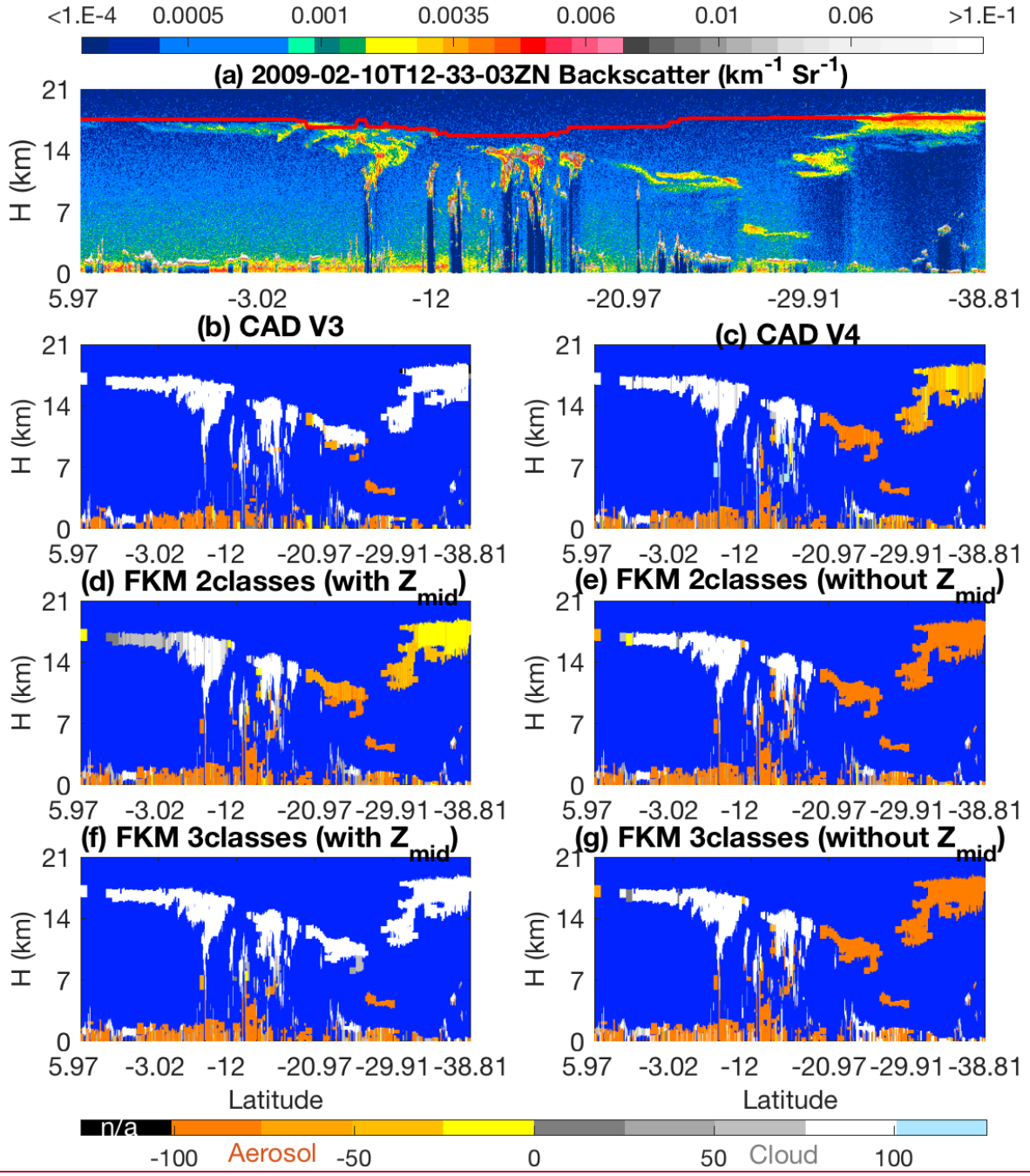


Figure 9: top row shows 532 nm attenuated backscatter coefficients for (a) measurements acquired on 10 February 2009 showing smoke injected into the upper troposphere and lower stratosphere by the Black Saturday fires in Australia. The solid line extending across (f) at altitudes between ~7 km and ~~(f) measurements acquired on~~ ~8.5 ~~October 2008 showing a layer of~~

~~volcanic ash from~~km shows the ~~eruption of Kasatochi.~~approximate tropopause altitude. The rows below show the CAD results reported by ~~four~~six different algorithms; the V3 operational CAD (~~panels~~ b ~~and g~~), the V4 operational CAD (~~panels~~ c ~~and h~~), the 2-class CAD_{FKM} ~~(panels~~ with z_{mid} (d) and ~~i~~without z_{mid} (e), and the 3-class CAD_{FKM} ~~(panels~~ with z_{mid} (e) and without z_{mid} (j).

c. ~~Tropospheric~~ Volcanic ~~Ash~~ash

Figure ~~9~~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 ~~ejecta~~'aerosol plume' near the tropopause at ~17:30:18 UCT ~~(Figure 9, panels f j). The~~. COCA V3 ~~operational CAD algorithm~~ classified those layers with base altitudes above the tropopause as 'stratospheric features' (black regions in Fig. 10b), and misclassified a substantial portion of ~~this ash plume~~the lower, tropospheric layers as cloud~~, and those~~. Those segments that were correctly classified as aerosol were frequently assigned low CAD scores. In contrast, ~~the~~COCA V4 ~~CAD algorithm~~ and both versions of the CAD_{FKM} with z_{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.

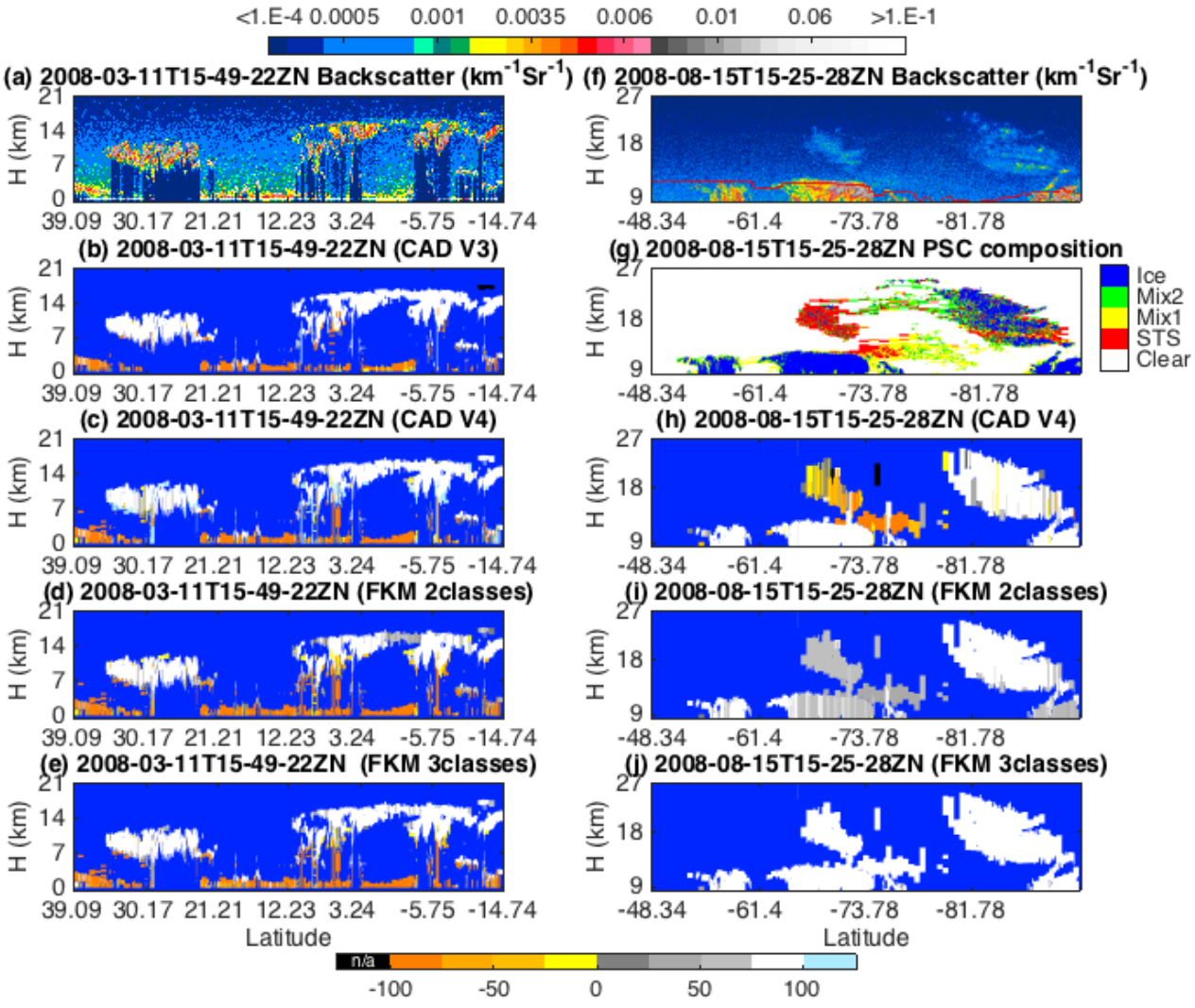
~~d. Fringes~~

~~The improved calibration coefficients in V4 facilitated the detection of optically thinner layers than were detected in V3 (Kar et al., 2018; Getzewich et al., 2018). A side effect of this improvement is an increased occurrence of optically thin layers detected along the tenuous edges of ice clouds. These layers, named "cirrus fringes", are detected at 20 km and 80 km horizontal resolutions, and occur along the sides of ice clouds or along their lower edges where overlying attenuation has substantially reduced the lidar signal. A "cirrus fringe amelioration" algorithm has been added in V4 as a CAD post processor which detects cirrus fringes that have been misclassified as aerosol and reclassifies them as cloud (Liu et al., 2018). These layers are given a special CAD score of 106. The misclassification of cirrus fringes as aerosols is basically due to their special nature. Fringes are optically thin and weakly scattering layers that occupy the transition zone between cirrus and clear sky. They are characterized by cirrus-like depolarization ratios coupled with lower color ratios, suggesting a reduction in ice crystal sizes possibly due to sublimation.~~

~~e. PSC~~

~~Polar Stratospheric Cloud (PSC) is the generic name for a class of clouds of several different compositions that all form in the winter polar stratosphere (Höpfner et al., 2009; Pitts et al., 2013). Prior to the V4 data release, all layers having base altitudes above the local tropopause were assigned to a generic class of 'stratospheric feature', and no further subtyping was done for these layers. In V4, however, the operational CAD algorithm now distinguishes between clouds and aerosols in the stratosphere~~

~~(Liu et al., 2018), and stratospheric aerosols are further evaluated to determine aerosol type (Kim et al., 2018). For PSCs, the CALIPSO project also produces a dedicated PSC data product that reports PSC presence in 5 km horizontal by 180 m vertical bins and classifies each bin according to a composition classification scheme described by Pitts et al. (2018). The compositions in the dedicated PSC product include water ice, supercooled ternary solutions (STS), and several mixtures of liquid droplets and nitric acid trihydrate (NAT). Figure 10 (panels f j) shows an example of PSC measurements from ~15:25:28 UCT on 15th August 2008. The V4 classifications agree well with the composition classifications developed by Pitts et al. (2009). High confidence CAD scores are given to those PSCs containing water ice particles, while low confidence CAD scores are given to low concentrations of liquid/NAT mixtures. The FKM classifications of the stratospheric features in this case do not agree with the V4 CAD or Pitts composition results. While there are some cases that agree well, these are relatively few and are not shown here. Even when the altitude dimension is excluded from the input parameters, classification of PSCs remains challenging for the FKM algorithm.~~


<1.E-4 0.0005 0.001 0.0035 0.006 0.01 0.06 >1.E-1
(a) 2008-03-11T15-49-22ZN Backscatter (km-1Sr-1)
(f) 2008-08-15T15-25-28ZN Backscatter (km-1Sr-1)
(b) 2008-03-11T15-49-22ZN (CAD V3)
(g) 2008-08-15T15-25-28ZN PSC composition
Ice
Mix2
Mix1
STS
Clear
(c) 2008-03-11T15-49-22ZN (CAD V4)
(h) 2008-08-15T15-25-28ZN (CAD V4)
(d) 2008-03-11T15-49-22ZN (FKM 2classes)
(i) 2008-08-15T15-25-28ZN (FKM 2classes)
(e) 2008-03-11T15-49-22ZN (FKM 3classes)
(j) 2008-08-15T15-25-28ZN (FKM 3classes)
H (km)
21
14
7
0
27
18
9
39.09 30.17 21.21 12.23 3.24 -5.75 -14.74
-48.34 -61.4 -73.78 -81.78
Latitude
n/a
-100 -50 0 50 100

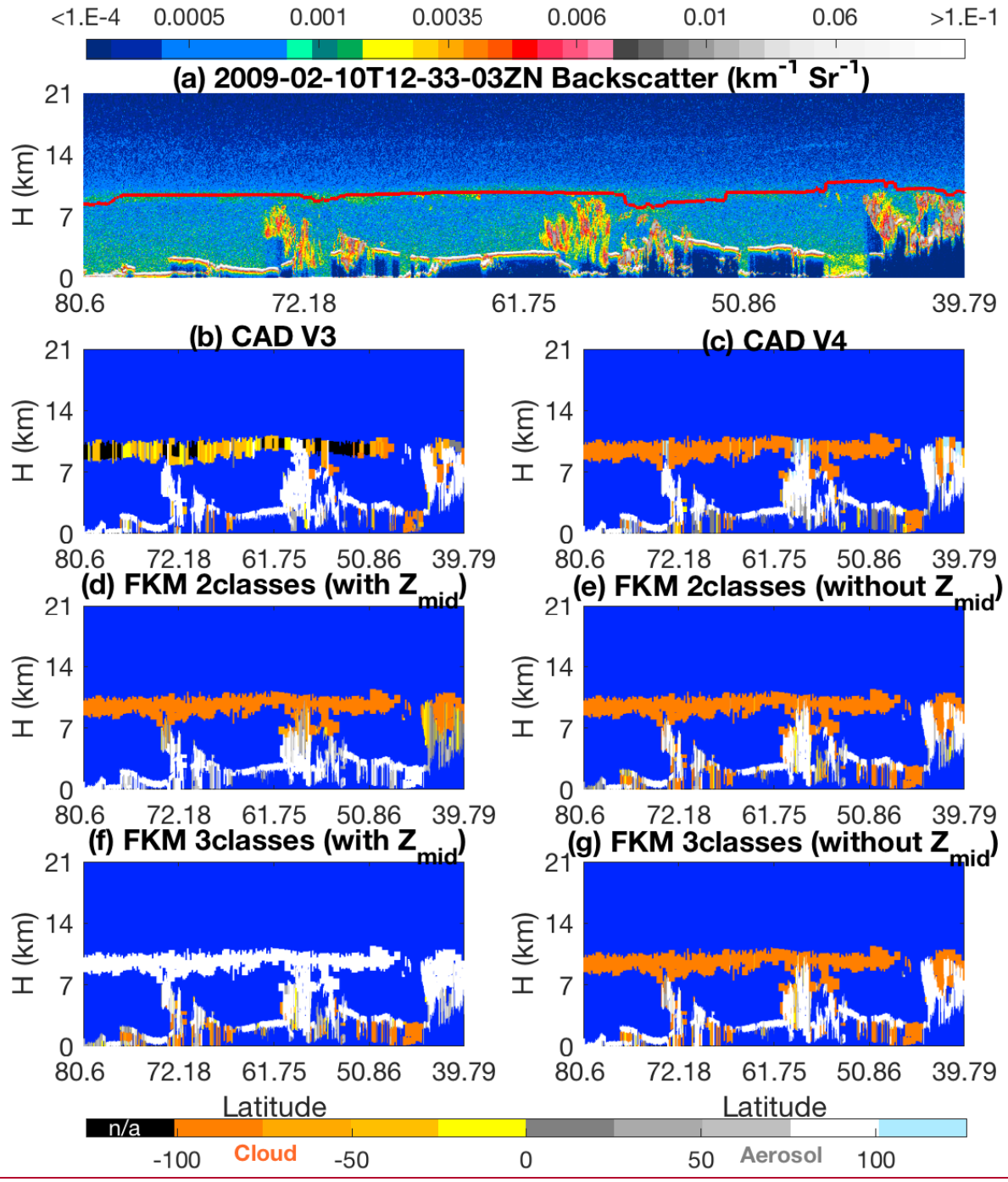


Figure 10: top row shows 532 nm attenuated backscatter coefficients for (~~panel a) extensive ice clouds~~a) measurements acquired on ~~10 July 2008 and (f) PSCs overlying tropospheric ice clouds on 15 August 2008. The second row shows (panel~~

~~b) V3 operational CAD results for the 10 July 2008 data and (panel g) the PSC classifications~~5 October 2008 showing a layer of volcanic ash from the eruption of Kasatochi. The solid line extending across (f) at altitudes between ~7 km and ~8.5 km shows the approximate tropopause altitude. The rows below show the CAD results reported ~~in~~by six different algorithms; the ~~dedicated CALIPSO PSC data product for the 15 August 2008 data. The three remaining rows show CAD results for the two scenes computed by~~ V3 operational CAD (b), the V4 operational CAD (~~panels~~ c ~~and h~~), the 2-class ~~FKM algorithm (panels~~ CAD_{FKM} with z_{mid} (d) and ~~i~~without z_{mid} (e), and the 3-class ~~FKM algorithm (panels e and j). The 10 July 2008 results are always in the left column and the 15 August 2008 results are always in the right column. For the PSC classifications in panel g, mix 1 and mix 2 (shown in yellow and green, respectively) represent different mixing ratios of STS and NAT.~~CAD_{FKM} with z_{mid} (e) and without z_{mid} (j).

~~4~~5 Discussion

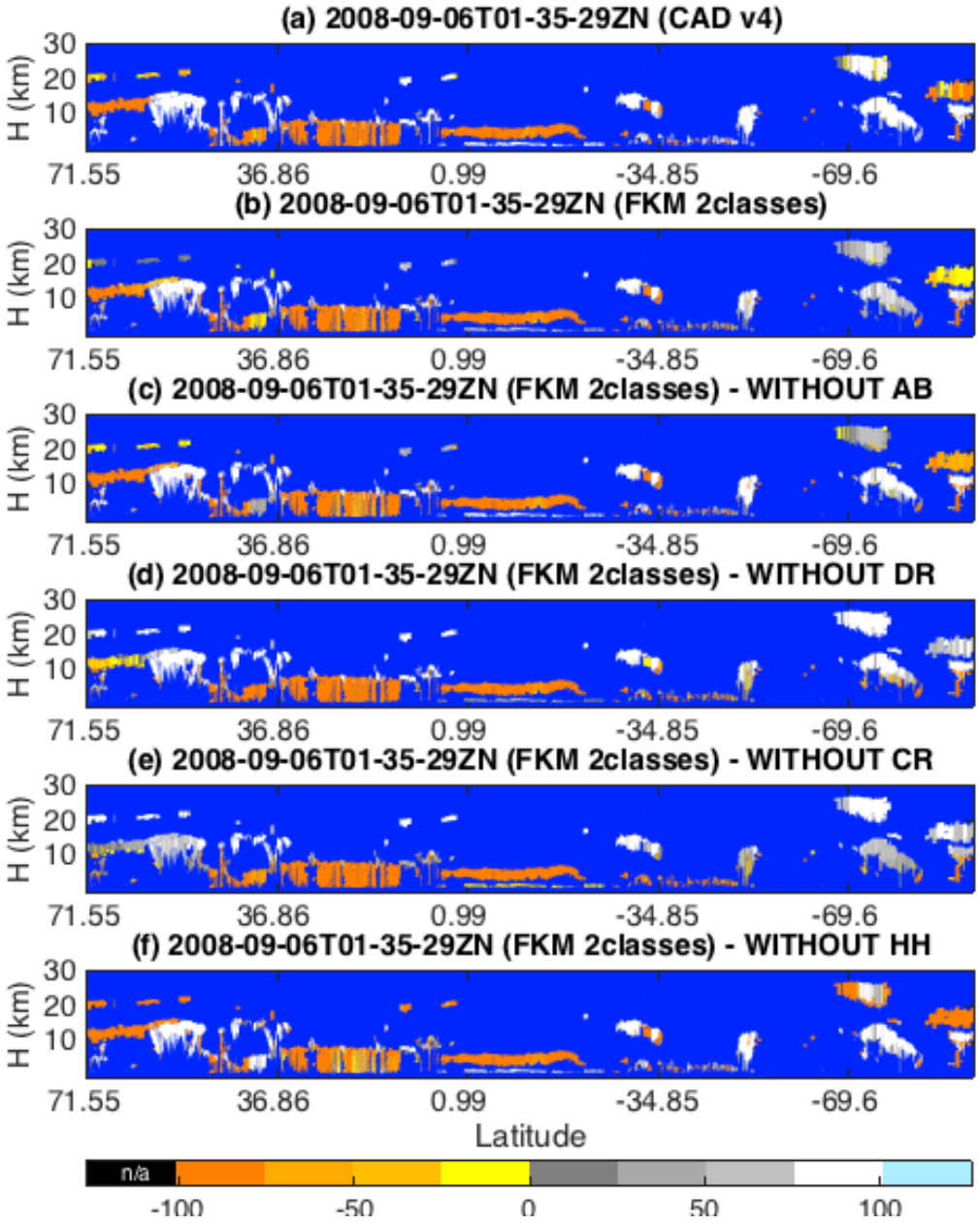
~~In~~ Section 5 compares FKM and COCA using statistical analyses and individual case studies. In this section ~~3~~, we ~~demonstrated that both the FKM and operational CAD algorithms generate very similar cloud and aerosol classifications. In this section, we discuss key parameter analysis, fuzzy linear discriminant analysis, principle component analysis and error propagation to~~ explore ~~several classification~~ 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 (~~section 4~~Sec. 5.1~~),~~); (b) how well are the classes separated ~~are the current classifications (section 4.2),~~(Sect. 5.2); (c) what are the essential measurements required for accurately discriminating between ~~cloud~~clouds and ~~aerosol (sections 4~~aerosols (Sect. 5.1 and ~~4~~Sect. 5.3~~),~~); and ~~how~~(d) what effects do ~~the~~ measurement uncertainties (~~or~~ noise) ~~impact~~have on the classifications (section 5.4~~.4)?)~~?

~~4~~5.1 Key parameter analysis

~~CALIOP's operational CAD algorithm was~~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 ~~CAD~~ PDFs. ~~Higher~~In general, higher dimension PDFs should ~~generally~~ 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 ~~easy~~relatively easy (though perhaps time-consuming) to add or remove one or multiple observational dimensions ~~and re-cluster without re-building new PDFs.~~(i.e., inputs) and the reinitiate the training/learning algorithm. (This highly desirable flexibility is, unfortunately, wholly absent in the strictly supervised learning regime incorporated into COCA.) If ~~one~~a dimension is added (or removed) and the new classifications are ~~similar (or~~

~~inferior)~~essentially identical to the old ones, the added (or removed) dimension does not provide significant ~~new~~ information ~~to~~in the classification ~~(i.e., it is not as important as~~processes. On the other ~~dimensions in the classification). If~~hand, if the CAD values are improved (or degraded) by adding ~~(~~or removing~~)~~ a dimension, this dimension actually ~~adds (or removes)~~contributes dispositive information in the determination of the classification, and hence is key to separating clouds from aerosols. ~~Moreover, by~~By using the FKM method, we can readily determine ~~how many dimensions~~which parameters are ~~enough~~required (and, importantly, which are non-essential) for the ~~cloud and aerosol classification according~~resulting classifications to ~~our required classification~~meet predetermined accuracy specifications, either in general or for a particular class (e.g., dust~~), and how much~~). We can also quantify the improvement (or degradation) that occurs ~~additional dimensions~~when specific parameters are either added or removed.

~~Re~~ 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 ~~are shown in Figure 11 and a summary of~~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 2-class, 4-parameter CAD_{FKM} results are quantified by the confusion matrices ~~are~~ shown in Table 3. From both the figure and the table we find that ~~without~~the cloud-aerosol partitioning obtained when any one dimension~~, most new clusters and their CAD_{FKM} values are unchanged compared~~ is omitted is reasonably similar to ~~classification including the dimension (more than 75% of cases stay same).~~the partitioning reported by COCA and by the 2-class, 4-parameter FKM algorithm. Both algorithms are most sensitive to the removal of χ' (75.0 % similarity for COCA and 77.4 % similarity for FKM), and least sensitive to the removal of $<\beta'_{532}>$ (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, ~~both V3 and~~ the 2-class 4-parameter FKM classifications have low CAD_{FKM} values. When either ~~mid-layer altitude~~ z_{mid} or ~~backscatter~~$<\beta'_{532}>$ is removed from the classification parameters, the CAD_{FKM} values actually improve. Without ~~color ratio~~χ' or ~~depolarization ratio~~δ_v, the CAD_{FKM} 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 case, as also explained later in section 4.3.~~. In this ~~case, both color ratio and backscatter~~example, the values of χ' and $<\beta'_{532}>$ measured in the water cloud can bias the resulting CAD_{FKM} values due to ~~uncertainties in the measurements related to the~~ the strong absorption at 532 nm ~~from~~within the overlying smoke ~~above the~~ layer. ~~The attenuated backscatter of~~ $<\beta'_{532}>$~~for~~ these water clouds decreases and gets closer to the backscatter magnitudes expected from classic aerosols (e.g., Figure 6a~~)~~), while ~~color ratio~~χ' increases ~~and gets~~ far ~~away from those~~beyond values typical of classic aerosols ~~(Figure 6c).~~. Moreover, when omitting z_{mid}, high altitude aerosols and ice clouds are more readily and correctly separated, as are low altitude aerosols and water clouds. χ' or δ_v are key in separating high altitude aerosols and clouds.


(a) 2008-09-06T01-35-29ZN (CAD v4)
(b) 2008-09-06T01-35-29ZN (FKM 2classes)
(c) 2008-09-06T01-35-29ZN (FKM 2classes) - WITHOUT AB
(d) 2008-09-06T01-35-29ZN (FKM 2classes) - WITHOUT DR
(e) 2008-09-06T01-35-29ZN (FKM 2classes) - WITHOUT CR
(f) 2008-09-06T01-35-29ZN (FKM 2classes) - WITHOUT HH
H (km)
30
20
10
71.55
36.86
0.99
-34.85
-69.6
Latitude
n/a
-100
-50
0
50
100

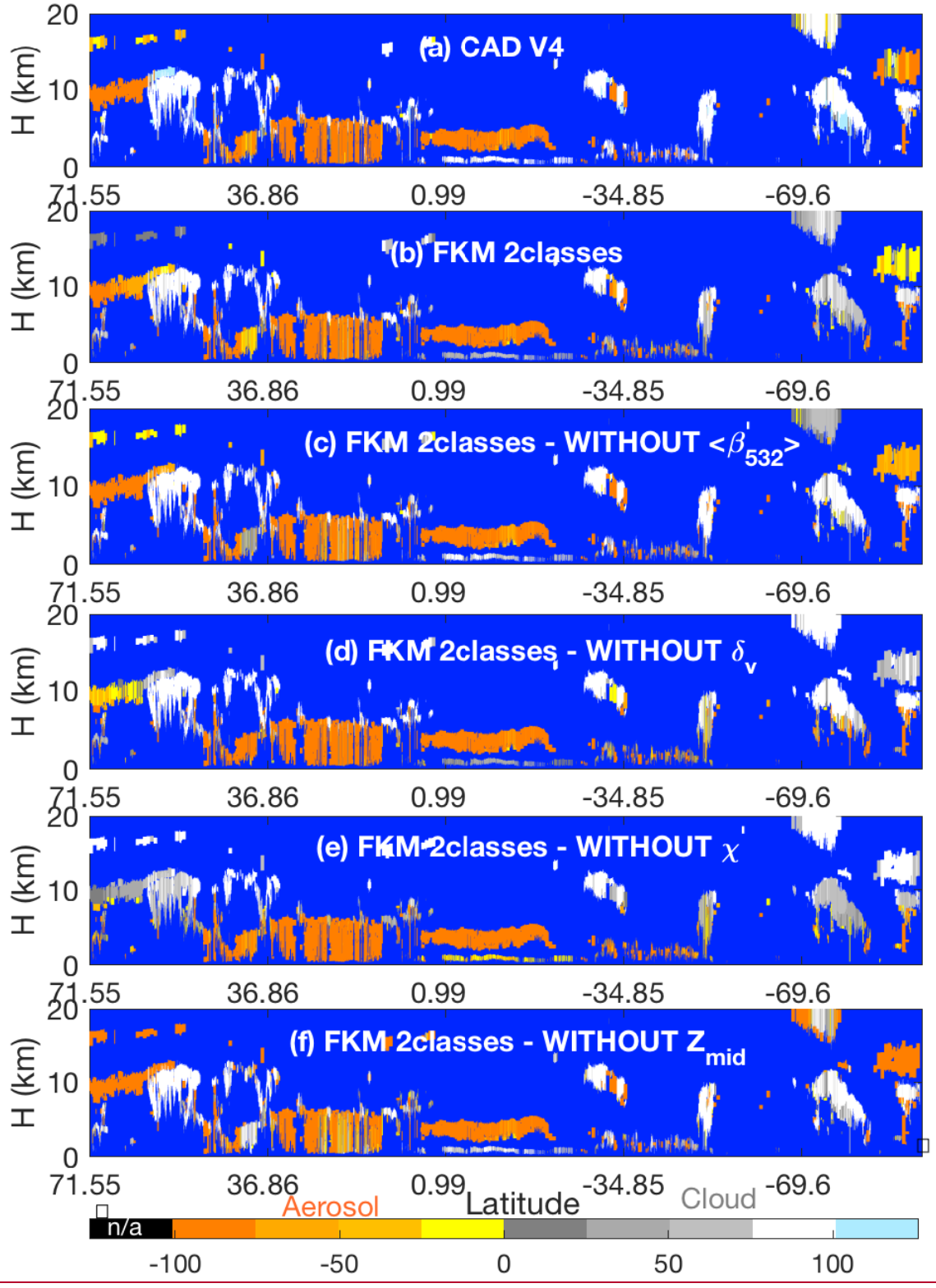

(a) CAD V4
(b) FKM 2classes
(c) FKM 2classes - WITHOUT <β'532>
(d) FKM 2classes - WITHOUT δv
(e) FKM 2classes - WITHOUT χ'
(f) FKM 2classes - WITHOUT Zmid
H (km)
20
10
0
71.55
36.86
0.99
-34.85
-69.6
Latitude
Aerosol
Cloud
n/a
-100
-50
0
50
100

Figure 11: CAD scores calculated using various techniques for an orbit segment on 6 September 2008 beginning at 01:35:29 UTC. The upper two rows show results from (a) the V4 operational CAD algorithm and (b) the 2-class FKM algorithm using all four standard inputs. The remaining rows show 2-class CAD_{FKM} results calculated when omitting one of the four standard inputs: (c) omits backscatter intensity~~, (<β′532>),~~ (d) omits depolarization ratios, (e) omits color ratios, and (f) omits mid-layer height.

Table 3: confusion matrices comparing COCA V4 and the CAD_{FKM} results shown in Figure 11; abbreviations as follows: C = cloud; A = aerosol; and T = total.

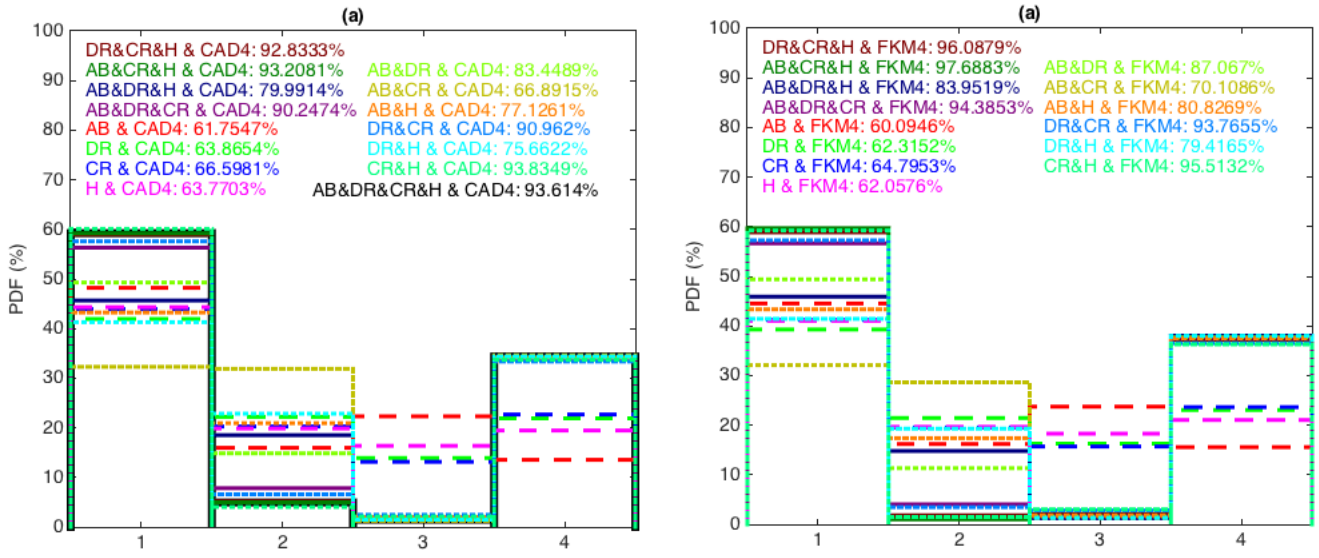
(%)		CAD_{FKM}			CAD_{FKM} (no ~~AB~~ <β′532>)			CAD_{FKM} (no ~~DR~~ δ_V)			CAD_{FKM} (no ~~CR~~ χ')			CAD_{FKM} (no ~~HH~~ Z_{mid})		
		C	A	T	C	A	T	C	A	T	C	A	T	C	A	T
~~CAD4~~COCA V4	C	45.1	3.3		45.6	2.8		43.7	1.7		40.1	11.4		44.9	3.5	
	A	7.3	44.3		7.4	44.2		12.7	38.3		13.6	34.9		12.3	39.6	
	T			89.4			89.8			82.0			75.0			84.5
CAD_{FKM}	C				56.9	2.6		52.3	7.2		49.6	10.0		55.6	3.2	
	A				4.3	36.1		3.7	36.8		12.6	27.9		8.4	32.1	
	T						93.1			89.1			77.4			87.8

~~Table 3: confusion matrices comparing V4 CAD and the CAD_{FKM} results shown in Figure 11; abbreviations as follows: AB = attenuated backscatter intensity; DR = depolarization ratio; CR = color ratio; H = mid layer altitude (height); C = cloud; A = aerosol; and T = total.~~

In addition to the ~~single half orbit tests~~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 CAD_{FKM} classifications using only a limited number of inputs (i.e., either 1, 2, or 3 of the CALIOP layer descriptors used in the standard ~~CAD_{FKM} classifications). The left panel of figure 12 shows the joint occurrence frequencies of these classifications and the V4 operational CAD classifications. Similarly, the right panel of figure 12 shows joint occurrence frequencies of the limited input classifications and the standard CAD_{FKM} using 4 dimensional (4-D) observations. Two distinct classes are considered, so that Figure 12 is, in effect, a linearized 2 x 2 confusion matrix, with values along the x axis representing each of the four different types of comparisons. Values plotted for x = 1 indicate the fraction of cases where both the limited input FKM and the V4 operational CAD algorithm (or the 4-D CAD_{FKM}) identified features as being in class 1. The results are color coded according to the number and type of the dimensions used in the limited input FKM method. Similarly, values plotted for x = 4 indicate the fraction of cases where both algorithms identified features~~

~~as being in class 2. x =1 and x = 4 correspond to the diagonal elements of the confusion matrix (i.e., M[1,1] and M[2,2]). The off-diagonal elements—i.e., M[1,2] and M[2,1]—are represented by, respectively, x = 2 and x = 3. For x = 2, the limited input FKM identifies the feature as belonging to class 1, while the V4 CAD (or 4-D CAD_{FKM}) identifies it as belonging to class 2. For x = 3 the assignments are reversed: the limited input FKM identifies the feature as belonging to class 2, whereas the V4 CAD (or 4-D CAD_{FKM}) identifies it as class 1~~2-class 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 2-class, 4-parameter FKM classifications.

~~From Figure 12b it is obvious that, even without any one observational dimension, more than 83% of classifications agree with results from four observational dimensions. Also, we notice that without color ratio, the agreement (only about 83%) between 3-D and 4-D FKM classifications is somewhat worse compared to results when omitting any of the other observations (agreement > 94%). We also find that when using only a single dimension of observations (e.g. backscatter or depolarization ratio or color ratio), FKM classification can only correctly separate the clouds and aerosols for about 60% of the cases compared to using 4-D CAD_{FKM} and/or the V4 operational CAD. This means that the additional 3 dimensions improve the cloud and aerosol discriminations by 30~40%. The more independent measurements that are used, the more accurate the classification can be. From Figure 12a, we again see that with 4-dimensional observations, the agreement between the V4 CAD and the FKM method is ~94%. This 6% differences may come from, for example, very thin, broken clouds, cloud fringes and dense aerosols that are inherently difficult to separate. When different combinations of observations are used in the classification, the disagreement between the two methods are different. V2 CAD algorithm used backscatter intensity, color ratio and mid layer altitude, which is the best combination of three independent parameters, showing only 0.4% fewer agreements compared to using four parameters. To further improve the CALIOP CAD algorithm, multiple investigations into additional combinations of observations or the use of weighted observations could be pursued in the future.~~



~~Figure 12: statistics of~~ Considering those comparisons where only one parameter is removed from the input data, it is clear that omitting χ' 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_{mid} rather than $<\beta'_{532}>$ or δ_v. The conclusions to be drawn from the single parameter classifications are similar to the 3-parameter case, though perhaps not as stark: using only χ' produced slightly better comparisons with both COCA V4 and the 2-class, 4-parameter CAD_{FKM} results than any of the other parameters. Given this demonstrated sensitivity to χ', it is perhaps not surprising that of the 2-parameter classifications, the combination of χ' and z_{mid} proves the most successful. The combination of χ' and δ_v also performed reasonably well relative to both COCA V4 and the 2-class, 4-parameter CAD_{FKM}. Unexpectedly, however, the combination of χ' and $<\beta'_{532}>$ performed very poorly relative to COCA V4, with only ~67 % of the classifications being identical.

In general, and as expected, the closest matches to the COCA V4 and the 2-class, 4-parameter CAD_{FKM} classifications are achieved by the 3-parameter classifications, with the single parameter classifications showing the poorest correspondences, and the 2-parameter rankings falling somewhere in between the 3-parameter and 1-parameter results. However, the performance of the most successful 2-parameter case (the combination of χ' and z_{mid}) was largely identical to that of the most successful 3-parameter case (the combination of $<\beta'_{532}>$, χ' and z_{mid}). In fact, relative to COCA V4, the 2-parameter classifications were identical slightly more often (93.8 % of all cases) than the 3-parameter classifications (93.2%). For the 2-class, 4-parameter FKM, the corresponding numbers rise to 95.5 % identity for the best performing 2-parameter classifications and 97.7 % identity for the best performing 3-parameter classifications. Both the COCA and FKM comparisons suggest that the addition of $<\beta'_{532}>$ 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.

Table 4: Statistics of joint occurrence frequency during January 2008, between 60°S and 60°N. The left panel (a) shows the joint occurrence of the V4 operational CAD classifications from 60°S to 60°N between and the FKM classifications based on limited input parameter sets (i.e., 1, 2, or 3 CALIOP measurements). The , as listed in the left column), the COCA V4 classifications (center column), and the 2-class, 4-parameter (2-C, 4-P) CAD_{FKM} classifications (right panel (b) shows the joint occurrence of the 4-D CAD_{FKM} classifications and the limited input FKM classifications. column).

-4

Occurrence Frequency (%)	V4 CAD	2-C, 4-P CAD_{FKM}
CAD_{FKM} (<β'_{532}>, δ_v, z_{mid}, χ')	93.61	-
CAD_{FKM} (<β'_{532}>, z_{mid}, χ')	93.21	97.69
CAD_{FKM} (δ_v, z_{mid}, χ')	92.83	96.09
CAD_{FKM} (<β'_{532}>, δ_v, χ')	90.25	94.39
CAD_{FKM} (<β'_{532}>, δ_v, z_{mid})	80.00	83.95
CAD_{FKM} (χ', z_{mid})	93.83	95.51
CAD_{FKM} (δ_v, χ')	90.96	93.77
CAD_{FKM} (<β'_{532}>, δ_v)	83.45	87.07
CAD_{FKM} (<β'_{532}>, z_{mid})	77.13	80.83
CAD_{FKM} (δ_v, z_{mid})	75.66	79.42
CAD_{FKM} (<β'_{532}>, χ')	66.89	70.11
CAD_{FKM} (χ')	66.60	64.80
CAD_{FKM} (δ_v)	63.87	62.32
CAD_{FKM} (z_{mid})	63.77	62.06
CAD_{FKM} (<β'_{532}>)	61.75	60.09

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_{il}, i, *data number* =1,..,n; l, *data dimension number* = 1,..p), the FKM classification returns a membership matrix M ($n \times k$ matrix, elements m_{ij}, i, *data number*= 1,..,n; j, *class number* = 1,..,k) and centroid matrix C ($k \times p$ matrix, elements c_{jl}, j, *class number* = 1,…,k; l, *data dimension number* = 1,…,p) where n is the number of data samples, p is the number of attributes/dimensions,

and k is the number of classes. The sums of squares and products (SSP) within-classes covariance matrix W_{lm} ($p \times p$ matrix, l/m, *data dimension number* = 1,…,p), also called the within-classes fuzzy scatter matrix (Bezdek, 1981), is given as

~~$W_{lm} = \sum_{j=1}^{k} \sum_{i=1}^{n} m_{ij}^{\phi} (x_{il} - c_{jl})(x_{im} - c_{jm}), \forall (l,m), l,m = 1, \dots, p$.~~ (16)

$$W_{lm} = \sum_{j=1}^{k} \sum_{i=1}^{n} m_{ij}^{\phi} (x_{il} - c_{jl})(x_{im} - c_{jm}), \forall (l,m), l,m = 1,2,\dots,p \quad . \tag{16}$$

The SSP between-classes covariance matrix B_{lm} ($p \times p$ matrix, l/m, *data dimension number* = 1,…,p) are given as

~~$B_{lm} = \sum_{j=1}^{k} (\sum_{i=1}^{n} m_{ij}^{\phi}) (c_{jl} - \bar{\bar{x}}_{l})(c_{jm} - \bar{\bar{x}}_{m}), \forall (l,m), l,m = 1, \dots, p$.~~ (17)

$$B_{lm} = \sum_{j=1}^{k} (\sum_{i=1}^{n} m_{ij}^{\phi}) (c_{jl} - \bar{\bar{x}}_{l})(c_{jm} - \bar{\bar{x}}_{m}), \forall (l,m), l,m = 1,2,\dots,p \quad . \tag{17}$$

The ratio of within-classes to the total SSP matrix is ~~the~~known as Wilks' lambda (Eq. 18, Wilks, 1932~~; Oh et al. 2005~~). Wilks' lambda for multi-dimensional observations is ~~a $p \times p$ matrix and~~ the determinant of the $p \times p$ matrix, which represents the geometric volume of this object in p dimensions, written as

~~$\Lambda = \frac{|W_{lm}|}{|W_{lm} + B_{lm}|}$.~~ (18)

$$\Lambda = \frac{\det(\mathrm{W})}{\det(\mathrm{W} + \mathrm{B})} \tag{18}$$

(Oh et al. 2005). Here we use Wilks' lambda (Λ) as a measure of the difference between classes~~, although a scalar cannot replace a vector for investigating different aspects of Wilks' lambda~~. The value Λ varies from 0 to 1, where 0 suggests that classes differ (within-classes SSP is smaller compared to between-classes SSP), and 1 suggests that all classes are the same. The magnitude of Wilks' Λ indicates how distinct and well-separated the classes are. Smaller values of Wilks' Λ indicate more distinct class separation with minimal between-class overlap thus the classification are more trustworthy and have higher confidence. Wilks' Λ 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' Λ for different ~~dimensional observations~~observational dimensions are calculated and summarized in Table ~~4~~5. For 4~~-~~dimensional (p=4) observations, Wilks' Λ could be as small as 0.21 for 2-class FKM and even

smaller (0.05) for 3-class FKM. This means that the classes generated by the FKM method are well separated, with clusters quite different from each other, ~~particularly~~and that the classes in 3-class FKM~~.~~ are much better separated (less overlap in the multi-dimensional observations) than the classes in 2-class FKM. For FKM 2-class, the value of Wilks' Λ is largest for ~~the mid-layer altitude dimension~~z_{mid}, indicating ~~a more overlaid classification compared~~that, relative to ~~classification~~the other individual parameters, clustering using ~~any other observational dimension.~~z_{mid} is less efficient at generating well-separated classes. The ~~reason for the~~ large value of Wilks' Λ ~~for the dimension altitude is~~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~~. These~~, and thus it is very difficult to separate these two classes using z_{mid} alone. The distinct altitude centers of ice and water clouds induce large within-classes SSP and hence large values of Wilks' Λ. ~~Following the values of~~For single parameter clustering, Wilks' Λ from ~~the dimension of altitude are the ones from color ratio and depolarization ratio, with the Wilks' Λ from backscatter intensity being~~$<\beta'_{532}>$ is the smallest~~.~~, followed by the values for δ_v and χ' . The value of Wilks' Λ from any combination ~~observations is~~ of observational dimensions lies between the maximum and minimum values ~~of those dimensions. For FKM~~ for the single parameter clustering. Wilks' Λ values for 3-class FKM are much smaller compared to the 2-class FKM values because z_{mid} can have an independent center for each class. For 3-class FKM analyses, ~~large~~the largest single parameter values of Wilks' Λ are produced by ~~depolarization, and~~ δ_v, followed by ~~altitude, color ratio and backscatter.~~z_{mid} and χ' . As with 2-class FKM, yields the smallest value.

Table 5: Wilks' lambda (Λ) for 2-class (center column) and 3-class (right column) FKM classifications using different observational dimensions (left column).

Input parameters	Λ, 2 classes	Λ, 3 classes
~~Backscatter intensity, depolarization ratio, color ratio, altitude~~$<\beta'_{532}>$, δ_v, χ', z_{mid}	0.21	0.048
~~Depolarization ratio, color ratio, altitude~~δ_v, χ', z_{mid}	0.20	0.060
~~Backscatter intensity, color ratio, altitude~~$<\beta'_{532}>$, χ', z_{mid}	0.20	0.060
~~Backscatter intensity, depolarization ratio, altitude~~$<\beta'_{532}>$, δ_v, z_{mid}	0.17	0.035
~~Backscatter intensity, depolarization ratio, color ratio~~$<\beta'_{532}>$, δ_v, χ'	0.14	0.030
~~Backscatter intensity and depolarization ratio~~$<\beta'_{532}>$, δ_v	0.12	0.025
~~Backscatter intensity and color ratio~~$<\beta'_{532}>$, χ'	0.14	0.039
~~Backscatter intensity and altitude~~$<\beta'_{532}>$, z_{mid}	0.14	0.052

Input parameters	Λ, 2 classes	Λ, 3 classes
~~Depolarization ratio and color ratio~~δ_v, χ'	0.13	0.043
~~Depolarization ratio and altitude~~δ_v, z_{mid}	0.20	0.056
~~Color ratio and altitude~~χ', z_{mid}	0.23	0.077
~~Backscatter intensity~~$<\beta'_{532}>$	0.08	0.030
~~Depolarization ratio~~δ_v	0.16	0.136
~~Color ratio~~χ'	0.18	0.053
~~Altitude~~z_{mid}	0.28	0.121

~~Table 4: Wilks' lambda (Λ) for 2-class (center column) and 3-class (rightmost column) FKM classifications using different dimensional observations (left column)~~

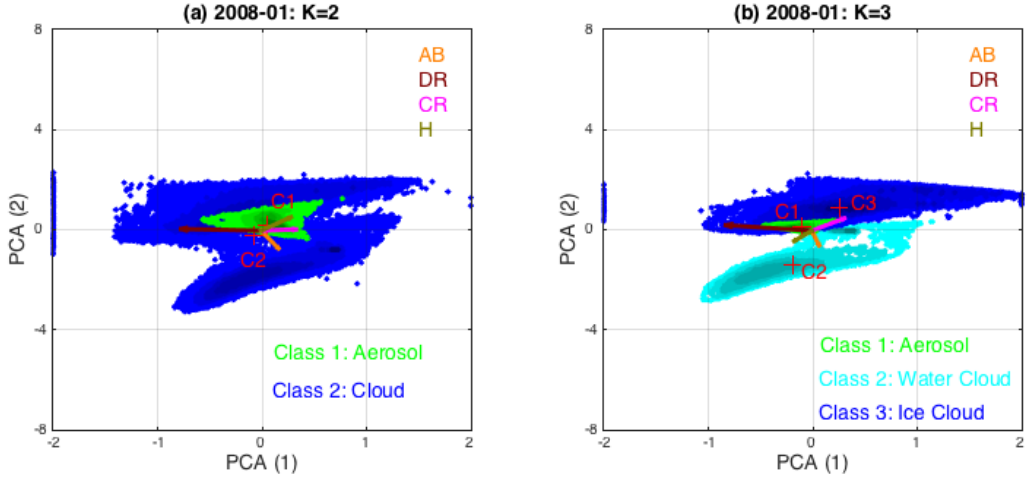
~~4~~

5.3 Principal Component Analysis

In this section we apply principal component analysis (PCA~~,~~; Wold et al. 1987) to the FKM ~~classifications~~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_f / B_f (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. ~~These functions, or canonical variants, are calculated from the eigenvalues and eigenvectors of matrix B_f/W_f (the ratio of within-class variance and between-class variance). Using this method will help us~~We also project 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 FKM 2 classes and 3 classes are shown in Figure ~~13~~12. The projection of vector lengths on PCA-1 and PCA-2 of different measurements (i.e., ~~backscatter intensity, depolarization ratio, color ratio~~$<\beta'_{532}>$, δ_v, and ~~mid-layer altitude~~χ', and z_{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). ~~Class~~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 2-class FKM breaks into 2 classes (ice cloud and water cloud) when applying 3-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 ~~color ratio~~χ' and ~~depolarization ratio~~δ_v contribute the most ~~on~~to PCA-1 (longer projections on the axis of PCA-1 in both subpanels) while ~~backscatter~~ and ~~height~~z_{mid} contribute more ~~on~~to PCA-2. Hence, ~~color ratio~~χ' and ~~depolarization~~δ_v are the driving components for the cloud ~~and~~ aerosol ~~classification~~separation. From figure ~~13b~~12b, we could also argue that ~~the backscatter and depolarization~~$\langle\beta'_{532}\rangle$ and δ_v are the driving factors in classifying water and ice clouds (projections of the vectors on ~~C1~~ C2-C3, namely the combined projection of PCA-1 and PCA-2, are longer), while ~~color ratio~~χ' and ~~altitude~~z_{mid} also contribute to the classification. ~~Altitude~~z_{mid}, and, to a greater extent, ~~color ratio~~χ' and ~~depolarization ratio~~δ_v are the driving factors that allow aerosols to be separated from ice clouds (projections of the vectors on C1-~~C3~~C2 are longer), whereas ~~backscatter intensity~~ and ~~color ratio~~χ' are the driving factors that separate water clouds from aerosols (projections of the vectors on ~~C2~~C1-C3 are longer). Comparing contributions of individual measurements to different classes, ~~mid layer altitude~~z_{mid} is most useful in helping discriminate aerosols and ice clouds, while simultaneously being the least useful in separating aerosols and water clouds ~~classification. Backscatter intensity~~. $\langle\beta'_{532}\rangle$ is the most useful parameter in ~~the~~distinguishing aerosols ~~vs.~~form water clouds ~~classification~~ and ~~the~~ water clouds ~~vs.~~from ice clouds ~~classification~~, and the least useful in ~~the~~differentiating between aerosols ~~vs.~~and ice clouds ~~classification. Depolarization~~. δ_v is most useful in distinguishing between water clouds and ice clouds and between aerosols and ice clouds, and the least useful in ~~the~~separating aerosols ~~vs.~~from water clouds ~~classifications.~~. These observations agree very well with earlier findings in ~~figures~~Fig. 6 and ~~12 and tables~~Tables 2, 3 and 4.



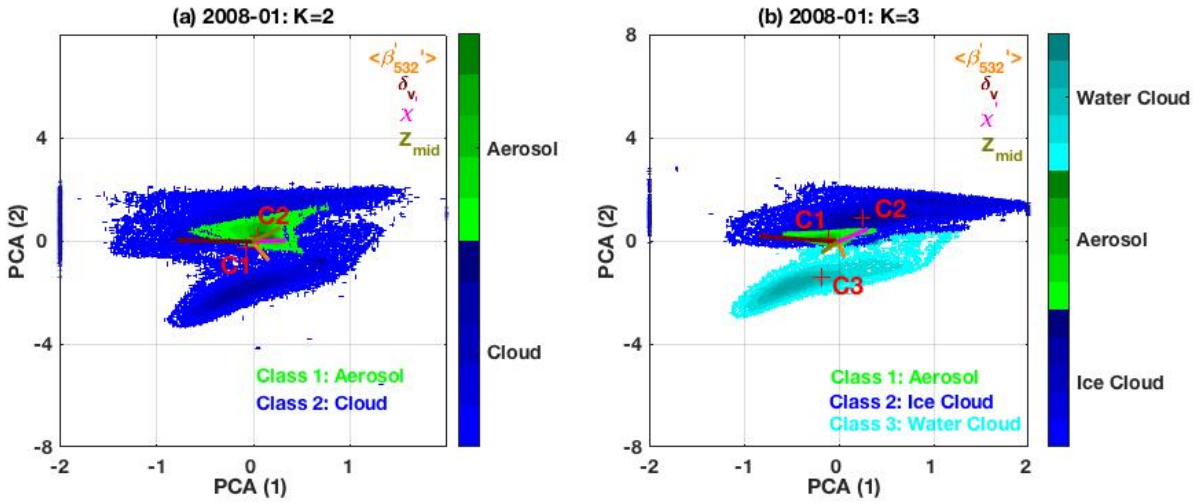


Figure ~~13~~12: principle component analysis of the FKM classifications for the January 2008 test data. PCA results for the 2-class CAD_{FKM} classifications are shown on the left (panel a), and the 3-class CAD_{FKM} classifications are shown on the right (panel b). In both figures, the green points are projections of aerosol data onto the PCA axes with their center located at red crosses labeled C1. Similarly, the blue and cyan (left panel only) points are projections of cloud data onto the PCA axes. In panel a, the blue points represent all clouds, while in panel b the blue points represent ice clouds and the cyan points represent water clouds. Higher sample number condensations are in darker colors while lower condensations are in lighter colors. Also shown in both panels are color-coded vectors representing each of the classification variables: backscatter intensity ~~(AB,~~($<\beta'_{532}>$, in orange), depolarization ratio (~~DR~~δ_v, in brown), color ratio (~~CR,~~χ', in magenta), and altitude (~~H~~z_{mid}, in olive). The projections of the variable vectors along the principal component axes indicate the degree to which each variable contributes to PCA1 and PCA2. -Variable vectors that are parallel to either PCA1 or PCA2 contribute essential information to that component, while vectors that are perpendicular do not contribute at all.

~~4~~5.4 Error propagation

~~In this section, we assess the impact~~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 measurement uncertainties on the FKM classifications. The observations from a nighttime granule acquired 6 September 2008 beginning at ~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, ~~biases~~errors are introduced into the cloud and aerosol classifications~~.~~ for this particular case. To simulate the measurement uncertainties, two different

methods are used. ~~In the~~The first ~~method,~~drew pseudo-random variables ~~were drawn~~ from Gaussian distributions having means equal to the various measured values and standard deviations between 10% and 200% of the means. As illustrated in Figure ~~14~~13, using this method allows us to quantify the effects of varying measurement errors on the FKM classification algorithm results. A sequence of Monte ~~Carlos~~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 ~~index~~indexes (CI, see section 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 ~~center~~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. 19 (Omar et al., 2005). These distances are normalized by the standard deviation of the distributions (C_{std}) of individual record distances from unperturbed center as

~~.~~

~~$$\delta d_e = \frac{|\overline{c_n - c_o}|}{c_{std}} \quad (19)$$~~

$$\delta d_\varepsilon = \frac{|c_n - c_0|}{c_{std}} \quad (19)$$

where |x| represents the L1 norm of x. Figure ~~14~~13 plots (a) the shifts in cluster centers ~~or (a)~~for each class, (b) the fraction of correct classifications, and (c) the revised confusion ~~index~~indexes as a function of relative uncertainties ranging from 10% to 200%. From Figure ~~14a~~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). Figures ~~14b~~13b and ~~14c~~13c show that when the uncertainties in the measurements are small (i.e. less than 10%), the errors in the classifications are also small (e.g., less than 2% in Figure ~~14b~~13b) with less overlaps between classes (e.g., small values of CI from 0.3 to 0.305 seen in Figure ~~14c~~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 ~~be stables~~tabilize 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 (~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 Figure 6.

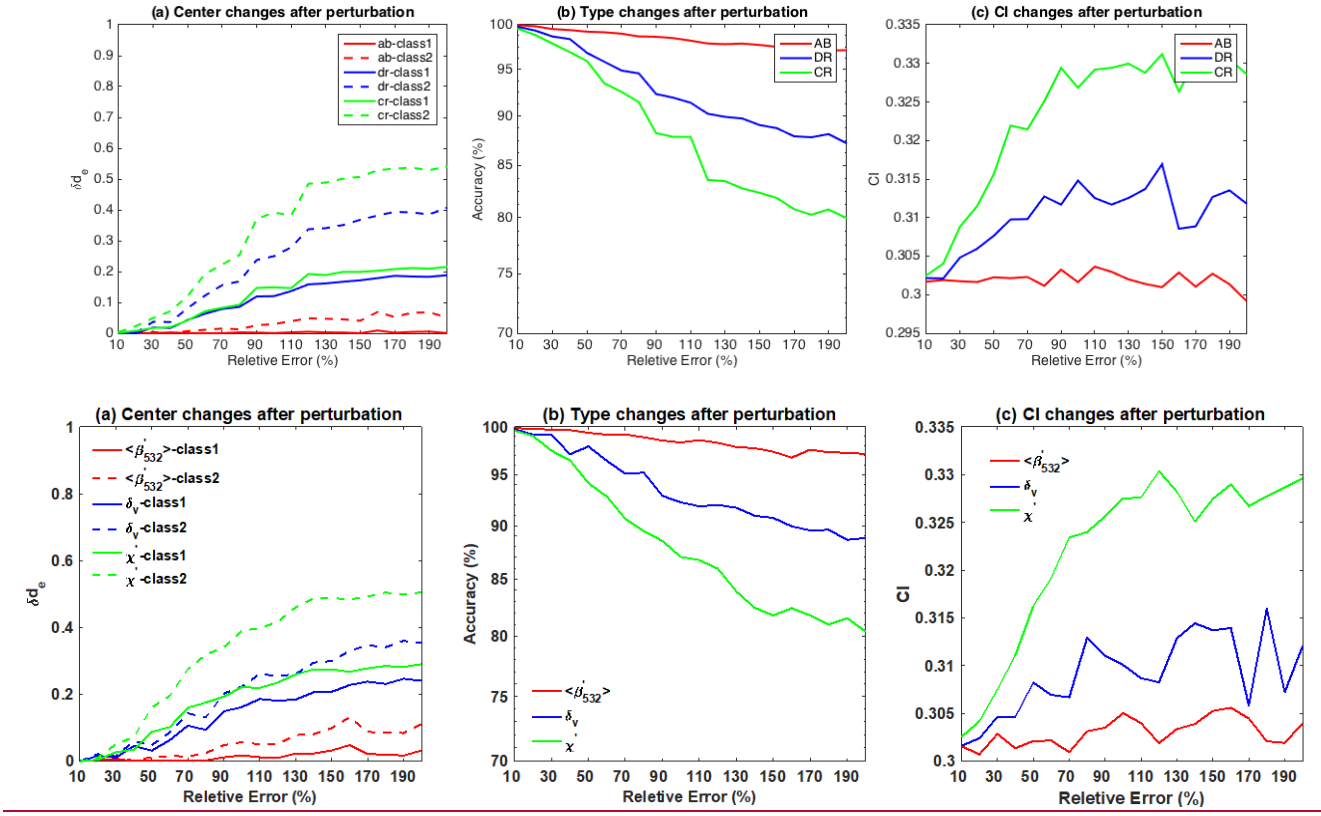
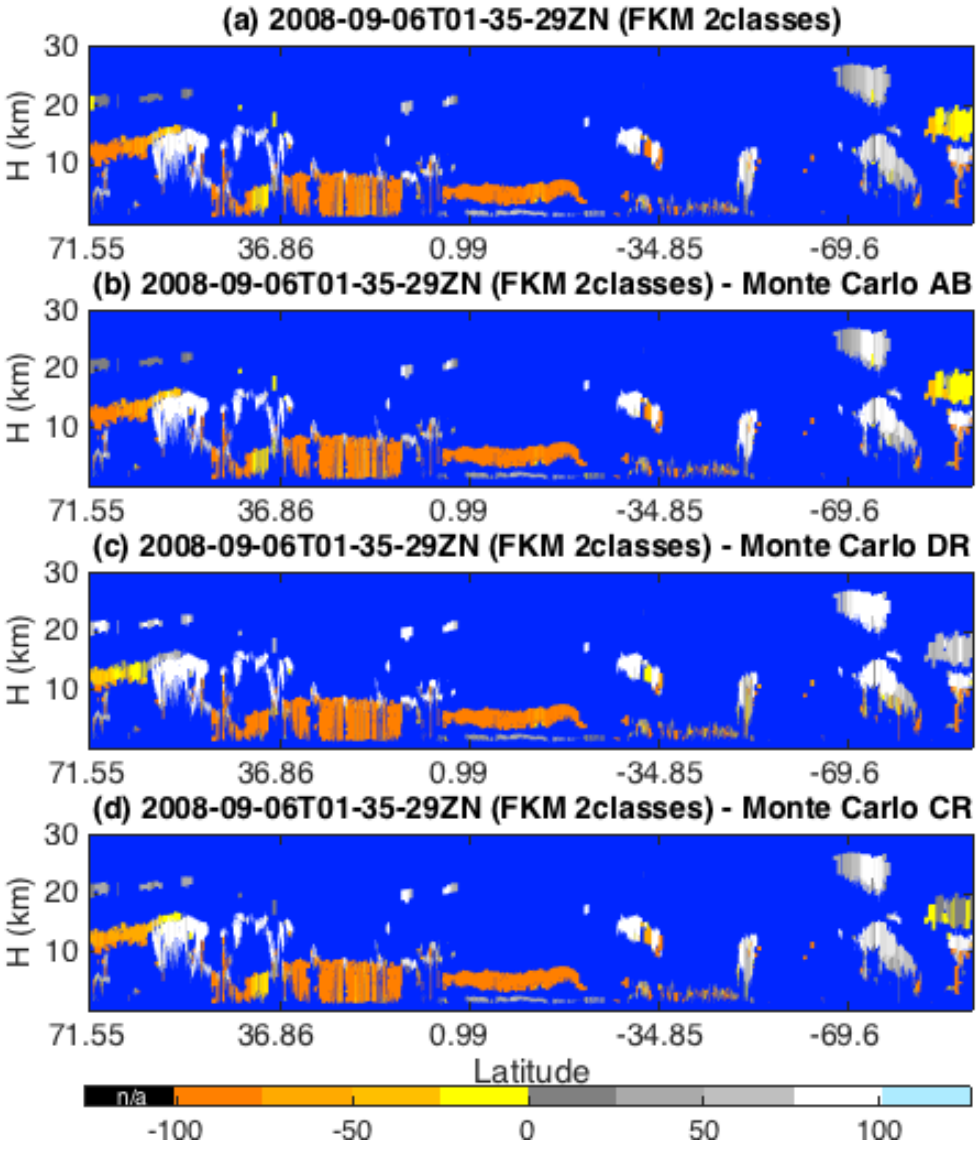


Figure ~~14~~13: classification changes as a function of errors in the input parameters.~~-~~ The left panel (a) shows shifts in cluster centers for each class; the center panel (b) shows the relative accuracy of the FKM classifications; and the right panel (c) shows changes in the cluster confusion indexes. Panels b and c show perturbations in the classifications due to uncertainties in attenuated backscatter intensity (~~AB,~~($<\beta'_{532}>$, in red), depolarization ratio (~~DR~~δ_v, in blue), and color ratio (~~CR,~~χ', in green).

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 construct 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 ~~15~~14 shows the results. Figure ~~15a~~14a shows the unperturbed results, while Figures ~~15b–15d~~14b–14d show CAD_{FKM} scores averaged over 100 perturbations of the test parameter. Figure ~~15b~~14b shows the results when the attenuated backscatter intensities are varied, Figure ~~15c~~14c shows the results when the depolarization ratios are varied, and Figure ~~15d~~14d shows the results when the color ratios are varied.

From Figure ~~15~~14 we find that the averaged CAD_{FKM} scores from the perturbed datasets do not differ markedly from the CAD_{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 (Figure ~~15b~~14b), we find that the perturbed and unperturbed classification results are identical more than 98% of the time. However, the CAD_{FKM} differences arising from perturbations to depolarization ratio and color ratio (Figures ~~15c~~14c and ~~15d~~14d, respectively) can be much larger. This finding is consistent with results shown earlier in Figure ~~14~~13.



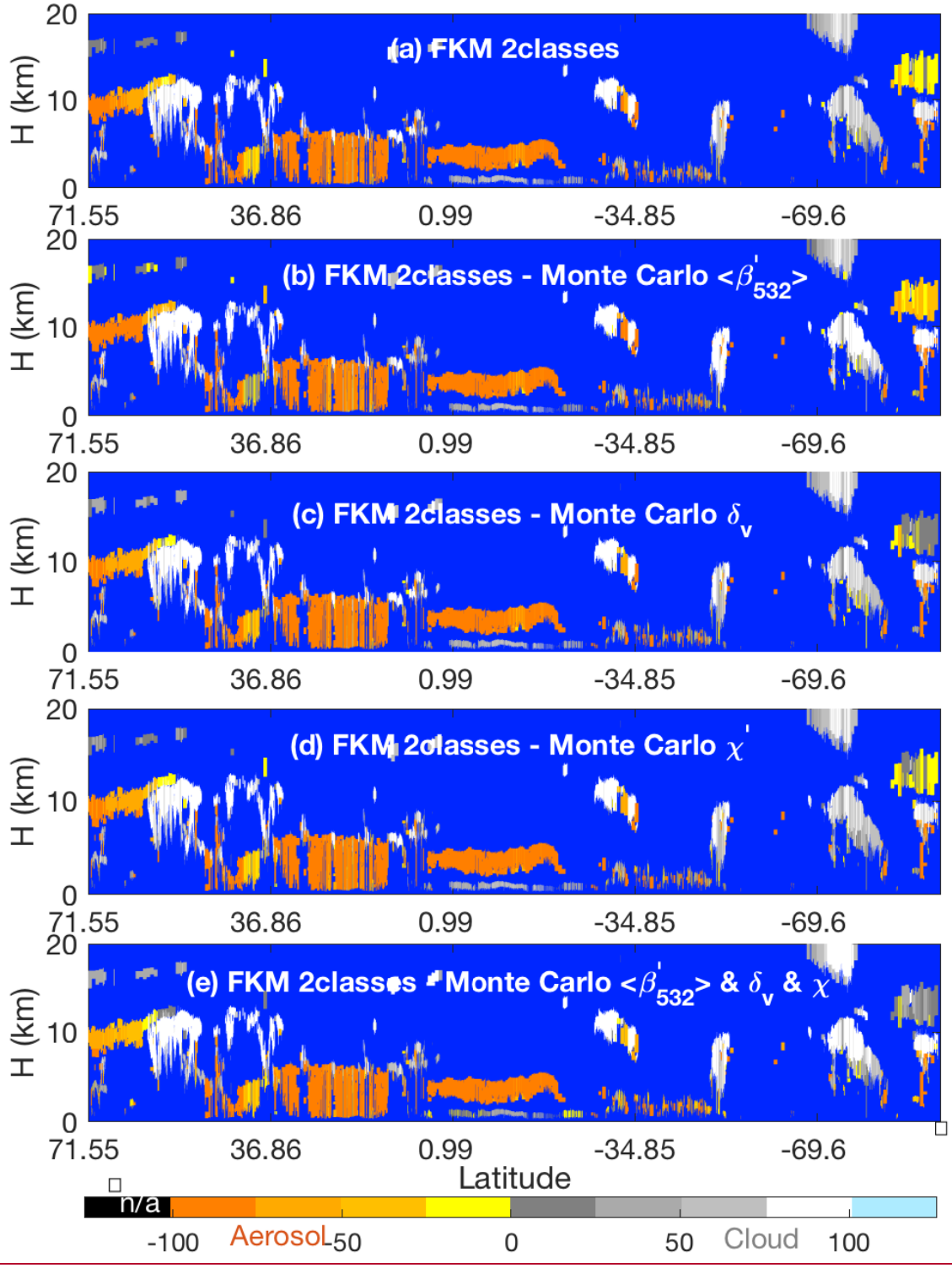

(a) FKM 2classes
(b) FKM 2classes - Monte Carlo <β'_{532}>
(c) FKM 2classes - Monte Carlo δ_v
(d) FKM 2classes - Monte Carlo χ'
(e) FKM 2classes - Monte Carlo <β'_{532}> & δ_v & χ'
H (km)
20
10
0
71.55
36.86
0.99
-34.85
-69.6
Latitude
n/a
-100
Aerosol
-50
0
50
Cloud
100

Figure ~~15~~14: CAD scores for the same orbit shown in Fig. 4 (06 Sep. 2008, 01:35:29 GMT). The uppermost panel (a) shows 2-class FKM results derived using unperturbed measurements. Panels b, c, and d show, respectively, 2-class FKM results derived using perturbed measurements of attenuated backscatter intensity ~~(AB),~~(<β'_{532}>), depolarization ratio (~~DR~~δ_v), and color ratio (~~CR).~~ χ'). Panel e shows 2-class FKM results derived when all three variables are perturbed independently.

Most often, the perturbed measurements only induce CAD_{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 Figure ~~15~~14, together with Figures 2, 11 and ~~13~~12, we find that this is a reasonable and even expected result. From Figures 11 and ~~15~~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 confidence CAD scores, both for the FKM method and the V4 operational method (Liu et al., 2018). A somewhat similar scenario can occur in the classification of high altitude aerosols, where high biases (i.e., measurement errors) in ~~depolarization ratios~~δ_v and ~~color ratios~~χ' can lead to the misclassification of aerosols as ice clouds.

~~5.~~6 Conclusions

In this paper we use the ~~Fuzzy K-Means~~fuzzy k-means (FKM) clustering algorithm to ~~validate~~evaluate the ~~performance of~~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 Observations (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.

~~To~~The comparison between two different classification techniques helps us assess the performance of the operational CAD algorithm~~, we have compared its outputs to those derived~~ from the same scenes ~~using an implementation of the FKM technique.~~ . The CALIOP operational CAD algorithm (COCA) is a supervised learning technique, in which classification decisions are tuned to match externally provided expert human judgements. ~~Unlike the operational CAD, the~~ 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 ~~use largely identical inputs~~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 ~~biases~~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 ~~CALIOP operational CAD algorithm~~ COCA are very closely replicated by the FKM method. ~~Our assessment of parameter PDFs shows that,~~Having a totally unsupervised learning algorithm "discover" the same patterns in ~~general, one half orbit of~~ the data ~~is sufficient to~~ that are reported by COCA strongly suggests that the COCA classifications represent ~~a much larger~~genuine data ~~set~~-driven differences in the ~~cloud aerosol classification process.~~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 ~~CALIOP's operational V4 CAD algorithm~~COCA and reported in the CALIOP V4 data products. Using a one-month test set, the 2-class and 3-class FKM classifications agreed with the V3 and V4 operational data products over 93 % of the time, and the 3-class FKM results agreed with the ~~V4 operational CAD~~COCA results in 94~95 % of all cases. This strong agreement between two independent methods provides convincing evidence that V4 ~~CAD~~ 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, ~~cirrus fringes,~~ 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, lofted dust layers and cirrus clouds occur in similar temperature regimes, and frequently have similar backscatter intensities ($<\beta'_{532}>$) and depolarization ratios~~.~~ (δ_v). The most critical criterion for distinguishing clouds from aerosols is color ratio~~,~~ (χ'), and the characteristic color ratios of dust and cirrus are reasonably distinct (~0.75 vs. ~1.01).~~-~~ However, the natural variability within each feature type is quite broad (e.g., ±0.25 for cirrus), and the measurements are very noisy, especially during daytime.

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~~, but the overall differences were not large.~~ . The primary differences are found by examining the results obtained for specific feature classes. The FKM classifications agree well with both the V3 ~~and V4~~ 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), as 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 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 ~~depolarization ratios~~δ_v and ~~color ratios;~~χ'; that ~~backscatter intensity~~$<\beta'_{532}>$ and ~~depolarization~~δ_v are the driving factors in classifying water and ice clouds; and that altitude~~, color ratio~~ (z_{mid}), χ' and ~~depolarization ratio~~δ_v are the key inputs that allow aerosols to be separated from ice clouds, while ~~backscatter intensity~~ 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.

~~While the FKM and official CAD classification methods both provide reliable discrimination between clouds and aerosols in the CALIOP data set, the FKM method is much more time consuming than the operational algorithm. On the other hand,~~ ~~the~~The flexibility of 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 is essentially ~~identical~~synthetic to that required by the CALIOP V4 operational algorithm, the two decision-making frameworks are independently derived and rely on very different mathematics (i.e., probabilities vs. 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 studied relied on the V4.10 CALIPSO level 2 vertical feature mask product (Vaughan et al., 2018b; NASA Langley Research Center Atmospheric Science Data Center; https://doi.org/10.5067/CALIOP/CALIPSO/LID_L2_VFM-Standard-V4-10; last access 22 May 2018) and the V4.10 CALIPSO level 2 5 km merged layer product (Vaughan et al., 2018b; NASA Langley Research Center Atmospheric Science Data Center; https://doi.org/10.5067/CALIOP/CALIPSO/LID_L2_05kmMLay-Standard-V4-10; last access 22 May 2018). 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

Competing interests: The authors declare that they have no conflicts of interest. Author Charles Trepte is a co-guest-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 affiliated with a conference.

References

Avery, M. A., Ryan, R., Getzewich, B., Vaughan, M., Winker, D., Hu, Y., and Trepte, C.: Impact of Near-Nadir Viewing Angles on CALIOP V4.1 Cloud Thermodynamic Phase Assignments, in preparation, 2018.

Bankert, R. L. and Solbrig, J. E.: Cluster Analysis of A-Train Data: Approximating the Vertical Cloud Structure of Oceanic Cloud Regimes, J. Appl. Meteor. Climatol., 54, 996–1008. doi:http://dx/doi.org/10.1175/JAMC-D-14-0227.1, 2015.

Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

Bezdek, J. C., Ehrlich, R., and Full, W.: FCM: the fuzzy c-means clustering algorithm, Comput. Geosci., 10, 191-203, 1984.

Burton, S. P., Ferrare, R. A., Hostetler, C. A., Hair, J. W., Rogers, R. R., Obland, M. D., Butler, C. F., Cook, A. L., Harper, D. B., and Froyd, K. D.: Aerosol classification using airborne High Spectral Resolution Lidar measurements - methodology and examples, Atmos. Meas. Tech., 5, 73-98, doi:10.5194/amt-5-73-2012, 2013.

Burrough P. A. and McDonnell R. A.: Principles of Geographic Information Systems, Oxford University Press, Oxford, 1998.

Burrough, P. A., Van Gaans, P. F. M., and MacMillan, R. A.: High-resolution landform classification using fuzzy K-means. Fuzzy Set, Syst., 113, 37–52, 2000.

Burrough ~~P. A. and McDonnell R. A.: Principles of Geographic Information Systems, Oxford University Press, Oxford, 1998~~, P. A., Wilson, J. P., van Gaans, P. F. M., and Hansen, A. J.: Fuzzy k-means classification of topo-climatic data as an aid to forest mapping in the Greater Yellowstone Area, USA, Landscape Ecology, 16, 523–546, https://doi.org/10.1023/A:1013167712622, 2001.

~~Chen, B., Huang, J., Minnis,~~Burton, S. P., ~~Hu, Y., Yi, Y., Liu, Z., Zhang, D~~Ferrare, R. A., Vaughan, M. A., Omar, A. H., Rogers, R. R., Hostetler, C. A., and ~~Wang, X.: Detection of dust aerosol by combining~~Hair, J. W.: Aerosol classification from airborne HSRL and comparisons with the CALIPSO ~~active lidar and passive IIR measurements~~vertical feature mask, Atmos. ~~Chem. Phys., 10, 4241-4251~~Meas. Tech., 6, 1397-1412, https://doi.org/10.5194/~~acp-10-4241-2010, 2010~~amt-6-1397-2013, 2013.

Chepfer, H., Bony, S., Winker, D., Chiriaco, M., Dufresne, J.-L., Sèze, G.: Use of CALIPSO lidar observations to evaluate the cloudiness simulated by a climate model, Geophys. Res. Lett~~.~~., 35, L15704, DOI: 10.1029/2008GL034207, 2008.

de Laat, A. T. J., Stein Zweers, D. C., Boers, R., and Tuinder, O. N. E.: A solar escalator: Observational evidence of the self-

lifting of smoke and aerosols by absorption of solar radiation in the February 2009 Australian Black Saturday plume, J. Geophys. Res., 117, D04204, https://doi.org/10.1029/2011JD017016, 2012.

DeMott, P. J., Prenni, A. J., Liu, X., Kreidenweis, S. M., Petters, M. D., Twohy, C. H., Richardson, M. S., Eidhammer, T., and Rogers, D. C.: Predicting global atmospheric ice nuclei distributions and their impacts on climate, PNAS, 107, 11217-11222, doi:10.1073/pnas.0910818107, 2010.

~~Di Pierro, M., Jaeglé, L., and Anderson, T. L.: Satellite observations of aerosol transport from East Asia to the Arctic: three case studies, Atmos. Chem. Phys., 11, 2225-2243, doi:10.5194/acp-11-2225-2011, 2011.~~

Fisher, R. A: The use of multiple measurements in taxonomic problems, Ann. Eugen. (Lond), 7, 179-188, 1936.

Gharibzadeh, M., Alam, K., Abedini, Y., Bidokhti, A. A., Masoumi, A., and Bibi, H.: Characterization of aerosol optical properties using multiple clustering techniques over Zanjan, Iran, during 2010–2013, Appl. Opt., 57, 2881–2889, doi:10.1364/AO.57.002881, 2018.

Getzewich, B. J., Vaughan, M. A., Hunt, W. H., Avery, M. A., Powell, K. A., Tackett, J. L., Winker, D. M., Kar, J., Lee, K.-P., and Toth, T~~., and Powell, K.~~: CALIPSO Lidar Calibration at 532– nm: Version 4 Daytime Algorithm, ~~in preparation~~Atmos. Meas. Tech., 11, 6309–6326, https://doi.org/10.5194/amt-11-6309-2018, 2018.

Gorsevski , P. V., Gessler, P. E. and Jankowski P.: Integrating a fuzzy K-means classification and a Bayesian approach for spatial prediction of landslide hazard, J. Geograph. Syst., 5, 223-251, 2003.

Harr, P. A., and R. L. Elsberry: Large-Scale Circulation Variability over the Tropical Western North Pacific. Part I: Spatial Patterns and Tropical Cyclone Characteristics, Monthly Weather Review, 123, 1225–1246, https://doi.org/10.1175/1520-0493(1995)123<1225:LSCVOT>2.0.CO;2, 1995.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-means Clustering Algorithm, J. Roy. Statist. Soc. Ser. B, 28, 100–108, 1979.

~~Höpfner~~Hlavka, D., Yorks, J., Young, S., Vaughan, M., Kuehn, R., McGill, M~~. Pitts,~~., and ~~L. Poole, 2009: "Comparison between~~Rodier, S.,: Airborne Validation of Cirrus Cloud Properties Derived from CALIPSO ~~and MIPAS observations of polar stratospheric clouds", J.~~Lidar Measurements: Optical Properties, J. Geophys. Res., ~~114, D00H05~~117, D09207, doi:10.1029/~~2009JD012114~~2011JD017053, 2012.

Hu, Y., Winker, D., Vaughan, M., Lin, B., Omar, A., Trepte, C., Flittner, D., Yang, P., Nasiri, S. L., Baum, B., Holz, R., Sun,

W., Liu, Z., Wang, Z., Young, S., Stamnes, K., Huang, J., Kuehn, R.: CALIPSO/CALIOP cloud phase discrimination algorithm, J. Atmos. Oceanic Technol., 26, 2293-2309, 2009.

~~Jin, Y., Kai, K., Okamoto, H., and Hagihara, Y.: Improvement of CALIOP cloud masking algorithms for better estimation of dust extinction profiles, J. Meteorol. Soc., 92, 433-455, 2014.~~

Jabari, S., and Zhang, Y.: Very High Resolution Satellite Image Classification Using Fuzzy Rule-Based Systems, Algorithms, 6, 762-781; doi:10.3390/a6040762, 2013.

Kar, J., Vaughan, M. A., Lee, K.-P., Tackett, J. L., Avery, M. A., Garnier, A., Getzewich, B. J., Hunt, W. H., Josset, D., Liu, Z., Lucker, P. L., Magill, B., Omar, A. H., Pelon, J., Rogers, R. R., Toth, T. D., Trepte, C. R., Vernier, J.-P., Winker, D. M., and Young, S. A.: CALIPSO lidar calibration at 532 nm: version 4 nighttime algorithm, Atmos. Meas. Tech., 11, 1459–1479, doi:10.5194/amt-11-1459-2018, 2018.

Key, J. R., Maslanik, J. A. and Barry, R. G.: Cloud classification from satellite data using a fuzzy sets algorithm: a polar example. International Journal of Remote Sensing, 10, 1823-1842, 1989.

Kim, M.-H., ~~Omar[2]~~Omar, A. H., Tackett, J. L., Vaughan, M. A., Winker, D. M., Trepte, C. R., Hu, Y., Liu, Z., Poole, L. R., Pitts, M. C., Kar, J., and Magill, B. E.: ~~Updates of the Version 4 CALIPSO Level 2 Aerosol Products, in preparation~~The CALIPSO version 4 automated aerosol classification and lidar ratio selection algorithm, Atmos. Meas. Tech., 11, 6107-6135, https://doi.org/10.5194/amt-11-6107-2018, 2018.

Konsta, D., Chepfer, H., and Dufresne, J. L.: Evaluation of Cloud Description in General Circulation Models Using A-Train Observations, Advances in Meteorology, Climatology and Atmospheric Physics. Springer Atmospheric Sciences. Springer, Berlin, Heidelberg, 2013.

Koren, I., Remer, L. A., Kaufman, Y. J., Rudich, Y., and Martins, J. V.: On the twilight zone between clouds and aerosols, Geophys. Res. Lett., 34, L08805, https://doi.org/10.1029/2007GL029253, 2007.

Kubat, M., Holte, R.C. and Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images, Machine Learning, vol. 30, nos. 2/3, pp. 195-215, 1998.

Liu, Z., Vaughan, M. A., Winker, D. M., Hostetler, C. A., Poole, L. R., Hlavka, D., Hart, W. and McGill, M.: Use of probability distribution functions for discriminating between cloud and aerosol in lidar backscatter data, J. Geophys. Res., 109, D15202, doi:10.1029/2004JD004732, 2004.

Liu, Z., Vaughan, M., Winker, D., Kittaka, C., Getzewich, B., Kuehn, R., Omar, A., Powell, K., Trepte, C. and Hostetler C.: The CALIPSO Lidar Cloud and Aerosol Discrimination: Version 2 Algorithm and Initial Assessment of Performance, J. Atmos. Oceanic Technol., 26, 1198–1213, doi:10.1175/2009JTECHA1229.1, 2009.

Liu, Z., Kar, J., Zeng, S., Tackett, J., Vaughan, M., Avery, M., Pelon, J., Getzewich, B., Lee, K.-P., Magill, B., Omar, A., Lucker, P., Trepte, C., and Winker, D.: Discriminating Between Clouds and Aerosols in the CALIOP Version 4.1 Data Products, ~~in preparation~~Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-190, in review, 2018.

Luo, Z. J., Anderson, R. C., Rossow, W. B., and Takahashi, H.: Tropical cloud and precipitation regimes as seen from near-simultaneous TRMM, CloudSat, and CALIPSO observations and comparison with ISCCP, J. Geophys. Res., 122, 5988–6003, doi:10.1002/2017JD026569, 2017.

Mahalanobis, P. C.: On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India, 2, 49–55, 1936.

~~Minasny, B., and McBratney, A. B.: FuzME version 3.0, Australian Centre for Precision Agriculture, The University of Sydney, Australia, 2002.~~

Mamouri, R. E., Amiridis, V., Papayannis, A., Giannakaki, E., Tsaknakis, G., and Balis, D. S.: Validation of CALIPSO space-borne-derived attenuated backscatter coefficient profiles using a ground-based lidar in Athens, Greece, Atmos. Meas. Tech., 2, 513-522, https://doi.org/10.5194/amt-2-513-2009, 2009.

McBratney, A. B., and de Gruijter, J. J.: A continuum approach to soil classification by modified fuzzy k-means with extragrades, European Journal of Soil Science, 43, 159–175, 1992.

McBratney, A.B., and Moore, A.W.: Application of fuzzy sets to climatic classification. Agricultural and Forest Meteorology, 35, 165–185, 1985

Metternicht, G.: Change detection assessment using fuzzy sets and remotely sensed data: an application of topographic map revision, ISPRS Journal of Photogrammetry & Remote Sensing, 54, 221–233, https://doi.org/10.1016/S0924-2716(99)00023-4, 1999.

Minasny, B., and McBratney, A. B.: FuzME version 3.0, Australian Centre for Precision Agriculture, The University of Sydney, Australia, 2002.

Mielonen, 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 AERONET inversion data, Geophys. Res. Lett., 36, L18804, doi:10.1029/2009GL039609, 2009

Mioche, G., Josset, D., Gayet, J.-F., Pelon, J., Garnier, A., Minikin, A., and Schwarzenboeck, A.: Validation of the CALIPSO/CALIOP extinction coefficients from in situ observations in mid-latitude cirrus clouds during CIRCLE-2 experiment, J. Geophys. Res., 115, D00H25, doi:10.1029/2009JD012376, 2010.

Mona, L., Pappalardo, G., Amodeo, A., D'Amico, G., Madonna, F., Boselli, A., Giunta, A., Russo, F., and Cuomo, V.: One year of CNR-IMAA multi-wavelength Raman lidar measurements in coincidence with CALIPSO overpasses: Level 1 products comparison, Atmos. Chem. Phys., 9, 7213-7228, https://doi.org/10.5194/acp-9-7213-2009, 2009.

Nam, C. C. W. and Quaas, J.: Evaluation of Clouds and Precipitation in the ECHAM5 General Circulation Model Using CALIPSO and CloudSat Satellite Data, J. Climate, 2012.

Nock, R. and Nielsen, F.: On Weighting Clustering, IEEE Trans. Pattern Anal. Mach. Intell., 28, 1-13, 2006.

Odeh, I. O. A., McBratney, A. B., and Chittleborough, D. J.: Soil pattern recognition with fuzzy c-means: application to classification and soil landform interrelationships, Soil Sci. Soc. Am. J., 56, 505–516, 1992a.

Odeh, I. O. A., McBratney, A. B., and Chittleborough, D. J.: Fuzzy c-means and kriging for mapping soil as a continuous system. Soil Sci. Soc. Am. J., 56, 1848–1854, 1992b.

Olthof, I., and Latifovic, R.: Short-term response of arctic vegetation NDVI to temperature anomalies, International Journal of Remote Sensing, 28, 4823-4840, https://doi.org/10.1080/01431160701268996, 2007.

Oh, C., Tok, A., and Ritchie, S. G.: Real-Time Freeway Level of Service Using Inductive-Signature-Based Vehicle Reidentification System, IEEE Trans. Intell. Transp. Syst., 6, 138-146, 2005.

Omar, A. H., Won, J.-G., Winker, D. M., Yoon, S.-C., Dubovik, O., McCormick, M. P.: Development of global aerosol models using cluster analysis of Aerosol Robotic Network (AERONET) measurements, J. Geophys. Res. Atmos., 110(10D), doi:10.1029/2004JD004874, 2005.

Omar, A. H., Winker, D. M., Vaughan, M. A., Hu, Y., Trepte, C. R., Ferrare, R. A., Lee, K., Hostetler, C. A., Kittaka, C., Rogers, R. R., Kuehn, R. E., and Liu, Z.: The CALIPSO Automated Aerosol Classification and Lidar Ratio Selection Algorithm, J. Atmos. Oceanic Technol., 26, 1994 - 2014, doi:10.1175/2009JTECHA1231.1, 2009.

Omar, A. H., ~~Won, J. G.,~~ Winker, D. M., ~~Yoon, S. C., Dubovik, O., McCormick,~~Tackett, J. L., Giles, D. M~~. P.: Development of global.~~, Kar, J., Liu, Z., Vaughan, M. A., Powell, K. A., and Trepte, C. R.: CALIOP and AERONET aerosol ~~models using cluster analysis of Aerosol Robotic Network (AERONET) measurements, J.~~optical depth comparisons: One size fits none, J. Geophys. Res~~. Atmos., 110(10D), ~~., 118, 1–19, https://doi~~:.org/10.~~1029/2004JD004874, 2005~~1002/jgrd.50330, 2013.

~~Pitts, M. C., Poole, L. R., Lambert, A., Thomason, L. W.: An assessment of CALIOP polar stratospheric cloud composition classification, Atmos. Chem. Phys., 13, 2975-2988, doi:10.5194/acp-13-2975-2013, 2013.~~

~~Pitts, M. C., Poole, L. R.: Polar stratospheric cloud climatology based on CALIPSO spaceborne lidar measurements from 2006-2017, Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2018-234, in review, 2018.~~

Prata, A. T., Young, S. T., Siems, S. T., and Manton, M. J.: Lidar ratios of stratospheric volcanic ash and sulfate aerosols retrieved from CALIOP measurements, Atmos. Chem. Phys., 17, 8599-8618, 2017.

Redemann, J., Vaughan, M. A., Zhang, Q., Shinozuka, Y., Russell, P. B., Livingston, J. M., Kacenelenbogen, M., and Remer, L. A.: The comparison of MODIS-Aqua (C5) and CALIOP (V2 & V3) aerosol optical depth, Atmos. Chem. Phys., 12, 3025–3043, https://doi.org/10.5194/acp-12-3025-2012, 2012.

Rogers, R. R., Vaughan, M. A., Hostetler, C. A., Burton, S. P., Ferrare, R. A., Young, S. A., Hair, J. W., Obland, M. D., Harper, D. B., Cook, A. L., and Winker, D. M.: Looking Through the Haze: Evaluating the CALIPSO Level 2 Aerosol Optical Depth using Airborne High Spectral Resolution Lidar Data, Atmos. Meas. Tech., 7, 4317–4340, https://doi.org/10.5194/amt-7-4317-2014, 2014.

Roubens, M.: Fuzzy clustering algorithms and their cluster validity, Eur. J. Oper. Res., 10, 294-301, 1982.

Schuster, G. L., Vaughan, M., MacDonnell, D., Su, W., Winker, D., Dubovik, O., Lapyonok, T., and Trepte, C.: Comparison of CALIPSO aerosol optical depth retrievals to AERONET measurements, and a climatology for the lidar ratio of dust, Atmos. Chem. Phys., 12, 7431–7452, https://doi.org/10.5194/acp-12-7431-2012, 2012.

Stephens, G., D. Winker, J. Pelon, C. Trepte, D. Vane, C. Yuhas, T. L'Ecuyer and M. Lebsock, 2017: "CloudSat and CALIPSO within the A-Train: Ten years of actively observing the Earth system", B. Am. Meteorol. Soc., doi:10.1175/BAMS-D-16-0324.1, in press.

Stubenrauch, C. J., and Coauthors: Assessment of global cloud datasets from satellites: Project and database initiated by the GEWEX Radiation Panel. Bull. Amer. Meteor. Soc., 94, 1031–1049, https://doi.org/10.1175/BAMS-D-12-00117.1, 2013.

Tesche, M., Wandinger, U., Ansmann, A., Althausen, D., Müller, D., and Omar, A. H.: Ground-based validation of CALIPSO observations of dust and smoke in the Cape Verde region, J. Geophys. Res., 118, 2889–2902, doi:10.1002/jgrd.50248, 2013.

Thorsen, T. J., Fu, Q., and Comstock, J. M.: Comparison of the CALIPSO satellite and ground-based observations of cirrus clouds at the ARM TWP sites, *J. Geophys. Res.*, **116**, D21203, doi:10.1029/2011JD015970, 2011.

Trenberth, K. E., J. T. Fasullo, and J. Kiehl: Earth's global energy budget. Bull. Amer. Meteor. Soc., 90, 311–323, https://doi.org/10.1175/2008BAMS2634.1, 2009

Triantafilis, J., Odeh, I. O. A., Minasny, B., and McBratney, A. B.: Elucidation of physiographic and hydrogeological features of the lower Namoi valley using fuzzy K-means classification of EM34 data, Environ. Model. Softw.,18, 667–680, 2003.

Usman, B.: Satellite Imagery Land Cover Classification using K-Means Clustering Algorithm Computer Vision for Environmental Information Extraction. Elixir Comp. Sci. & Engg,. 63, 2013, 18671-18675, 2013

Vaughan, M., K. Powell, R. Kuehn, S. Young, D. Winker, C. Hostetler, W. Hunt, Z. Liu, M. McGill, B. Getzewich: Fully Automated Detection of Cloud and Aerosol Layers in the CALIPSO Lidar Measurements, J. Atmos. Oceanic Technol., 26, 2034–2050, doi:10.1175/2009JTECHA1228.1, 2009.

Vaughan, M., Garnier, A., Josset, D., Avery, M., Lee, K.-P., Liu, Z., Hunt, ~~B~~W., Pelon, J., Hu, Y., Burton, S., Hair, J., Tackett, J. L., Getzewich, B., Kar, J., and ~~Burton~~Rodier, S.: CALIPSO Lidar Calibration at 1064- nm: Version 4 Algorithm, ~~in preparation,~~ Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-303, in review, 2018a.

Vaughan, M., Pitts, M., Trepte, C., Winker, D., Detweiler, P., Garnier, A., Getzewich, B., Hunt, W., Lambeth, J., Lee, K.-P., Lucker, P., Murray, T., Rodier, S., Tremas, T., Bazureau, A. and Pelon, J.: Cloud-Aerosol LIDAR Infrared Pathfinder Satellite Observations (CALIPSO) data management system data products catalog, Release 4.40, NASA Langley Research Center Document PC-SCI-503, 2018b (available online at https://www-calipso.larc.nasa.gov/products/CALIPSO_DPC_Rev4x40.pdf; last access: 22 May 2018).

Wilks, S. S.: Certain generalizations in the analysis of variance, Biometrika, 24, 471-494, 1932.

Winker, D. M., R. H. Couch, and M. P. McCormick: An overview of LITE: NASA's Lidar In-space Technology Experiment, Proc. IEEE, 84, 164–180, doi:10.1109/5.482227, 1996.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y., Powell, K. A., Liu, Z., Hunt, W. H. and Young, S. A.: Overview of the

CALIPSO Mission and CALIOP Data Processing Algorithms, J. Atmos. Oceanic Technol., 26, 2310–2323, doi:10.1175/2009JTECHA1281.1, 2009.

Winker, D. M., J. Pelon, J. A. Coakley, Jr., S. A. Ackerman, R. J. Charlson, P. R. Colarco, P. Flamant, Q. Fu, R. Hoff, C. Kittaka, T. L. Kubar, H. LeTreut, M. P. McCormick, G. Megie, L. Poole, K. Powell, C. Trepte, M. A. Vaughan, B. A. Wielicki: "The CALIPSO Mission: A Global 3D View Of Aerosols And Clouds", Bull. Am. Meteorol. Soc., 91, 1211–1229, doi:10.1175/2010BAMS3009.1, 2010.

Wold, S., Esbensen, K., Geladi, P.: Principal component analysis, Chem. Intell. Lab. Syst., 2 37-52, doi:10.1016/0169-7439(87)80084-9, 1987.

Yorks, J., Hlavka, D., Vaughan, M., McGill, M., Hart, W., Rodier, S., and Kuehn, R.: Airborne Validation of Cirrus Cloud Properties Derived from CALIPSO Lidar Measurements: Spatial Properties, J. Geophys. Res., 116, D19207, doi:10.1029/2011JD015942, 2011.

Zhang, Y., Klein, S., Mace, G. G., and Boyle, J.: Cluster analysis of tropical clouds using CloudSat data, Geophys. Res. Lett., 34, L12813, doi:10.1029/2007GL029336, 2007.

Interactive comment on "Application of High-Dimensional Fuzzy K-means Cluster Analysis to CALIOP/CALIPSO Version 4.1 Cloud-Aerosol Discrimination" *by* Shan Zeng et al.

Anonymous Referee #1

This manuscript describes a cluster analysis technique applied to CALIPSO data products as an alternative to the standard cloud-aerosol discrimination (CAD) algorithm(s). The stated objective of this activity is to validate the CAD algorithm and better understand what is important in the classification process.

I believe the elements of a good paper are here, but the logical flow needs some significant work. I need to really dig to understand why you did this in the first place, and what specifically you learned when you were done. The authors used a large number of statistical and otherwise assessment techniques, but often gave little justification for the choices that were made during the assessment, and minimal description of the results. I think it would have been better to not describe every single analysis that was performed for this work, and instead focus on some of the most important and try for a better explanation of what you were doing, why you picked that technique, what you expected to find, and how the differences between expectations and results are significant to the CAD algorithms. The significance should be summarized both in the conclusions and the abstract. Before I start addressing specific portions of the manuscript, I have a few general questions and comments:

Thanks for reviewing the manuscript and giving many valuable suggestions. We re-worked the logical flow of our paper, and we have clarified our working plan in the introduction. We better interlinked the sections and subsections to make the flow of the manuscript clear and easy for readers and reviewers to follow. We have also added more descriptions and explanations of the results in section 3 and 4, concentrated on and more clearly identified especially significant results, and better summarized them in the abstract and the conclusion. Please check details in the paper.

1. It seems to me that hierarchical clustering, in one of its many forms, would be useful to CALIPSO data. Presumably, clouds are very different from aerosols, but ice and water clouds, while different, are much more similar to each other than to aerosols. This is not the topic of the paper, of course, by if one is to spend time looking into clustering algorithms as an alternative for CAD, why wasn't this considered?

   It seems we've done a poor job of stating our objectives for this paper, as it was never meant to be an exercise in "looking into clustering algorithms as an alternative for CAD". Instead, as we say in the conclusions of the original draft, the purpose of this study is "to assess the performance of the cloud-aerosol discrimination (CAD) algorithm used in the standard processing". Having now read the referee's comments, it's quite apparent that we failed to clearly enunciate our primary goal.

   Unsupervised clustering schemes are frequently used to find patterns in data. The trick is to find patterns that have meaningful interpretations for human data users. Having a totally unsupervised learning algorithm "discover" the same patterns in the data that we report when using our CAD PDF scheme gives us some confidence that our CAD classifications represent genuine data-driven differences in the CALIOP observations, and are thus largely free from artifacts that might be imposed by human misinterpretations or conceits. Furthermore,

interpreting our comparisons of the results obtained by the two methods is straightforward: the higher the correspondence between the "discovered" classes and our predetermined classes, the higher our confidence in the CALIOP operational CAD classifications.

2. At no point do you discuss the possibility of the FKM acting as a potential replacement for the standard CAD algorithms. Why not?

Once again it appears that we have failed to convey our goals for this study. We are not investigating potential replacements for our CAD algorithm. Instead we are looking for methods to validate its performance and characterize its reliability. We will, of course, replace the standard CAD algorithms if/when some other algorithm(s) – be it FKM or anything else – can be shown to yield ***consistent and demonstrable improvement*** in distinguishing between clouds and aerosols ***on a global scale***. But that day has not yet arrived. The take-away message from this paper is (or least should be!) that the FKM classifications are essentially similar to the operational V4 results, and thus confirm that our operational CAD algorithms are performing quite well. In the testing we've conducted to date, we see no evidence of superior performance by the FKM algorithm.

In the last paragraph, you mention that the "FKM method is much more time consuming than the operational algorithm." I don't understand this. Does this mean it is time consuming if one attempts to recreate the FKM for the entire dataset? I would think one would create the cluster centers with a subset of data and apply to the rest, which I couldn't imagine would be slower than the CAD approach.

We were regrettably imprecise in our comments about algorithm speeds. What's time consuming is training the FKM on a suitably large subset of the measurements. (In this regard, FKM is similar to our standard algorithm: building the CAD PDFs is also a time- and compute-intensive activity.) However, once the FKM training has been completed, the partitioning of features within any scene into clouds and aerosols should occur on time scales commensurate with (or perhaps faster than) our current CAD algorithm.

Other than speed, I see no discussion about why FKM would be less desirable than CAD. In some ways (less dependence on non-scene information like latitude), I would think it would be preferable.

Conceptually, both FKM and the existing algorithm are potentially good candidates for the CAD task. But practically speaking, our existing algorithm is the hands down winner: it's already highly optimized for use with CALIOP measurements and tightly integrated into the CALIOP software architecture, and its performance has been extensively documented (e.g., see Liu et al., 2009 and Liu et al., 2018).

Regarding the use of "non-scene information like latitude", recall that the CALIOP CAD algorithm is fundamentally a probability-driven technique. And because (for example) the probability of observing lofted dust plumes at 60° S is significantly smaller than at 45° N, latitude can contribute relevant information in distinguishing cloud (e.g., cirrus) from aerosol (e.g., Asian dust). On the other hand, using attributes like latitude and altitude can introduce classification artifacts into ***unsupervised*** machine learning techniques like fuzzy k-means and Kohonen self-organizing maps. These techniques are likely to be much more successful if the inputs can be restricted to the intrinsic properties of the atmospheric layers being measured (e.g., particulate depolarization ratios, lidar ratios, Ångström exponents, etc.). Unfortunately, these quantities are not directly

measured by elastic backscatter lidars like CALIOP and thus cannot be used in the classification phase of the analysis. Instead we have to make due with proxies like total attenuated backscatter color ratio, which, while readily measured, is also a rather awkward combination of intrinsic and extrinsic layer properties. Having multi-wavelength high spectral resolution lidar measurements would remedy the situation…but, sadly, it's likely to be many years before such a capable system is flown in space.

3. Much of your validation relies on a comparison of various versions of CAD and FKM. In some cases this is just for a few specific scenes. Once we make the leap of faith that those scenes are representative of all scenes, the fact remains that you don't know truth. So, is agreement between CAD and FKM the best metric when they could both be wrong? It seems the implicit assumption is that FKM could never reach CAD levels of correctness, but when if FKM does correctly identify the scene but CAD does not? In that case, 'agreement' isn't appropriate as a means to validate the results. One way of addressing this problem is to apply both algorithms to synthetically generated scenes, where one knows the 'truth' and can verify its identification.

   To say that "much of your validation relies on a comparison of various versions of CAD and FKM" is perhaps being too kind. A more realistic assessment might be that **ALL** of our validation currently relies on these comparisons. To the best of our knowledge, there are no published, observation-based validation studies of the CALIOP CAD results. This paper is our first attempt to evaluate the CALIOP cloud-aerosol discrimination problem within a different mathematical decision-making framework. (We note, though, that since our manuscript first appeared in the AMT discussion forum, another algorithm-comparison study using state vector machines has been published; see Brakhasi et al., 2018 at https://doi.org/10.1016/j.jag.2018.07.017.)

   While the use of synthetic data as an evaluation tool is generally an excellent and highly effective strategy, in the case of discriminating clouds from aerosols it's also especially hard to implement in a useful way. For ~90% of the cases, cloud and aerosol properties are very well separated and reliable classifications can be made using a single wavelength elastic backscatter lidar (e.g., CATS; see https://cats.gsfc.nasa.gov/media/docs/CATS_QS_L2O_Layer_3.00.pdf). In these cases, unambiguous synthetic data can be used to weed out those algorithms that are obviously deficient. But the remaining cases fall into the cloud-aerosol overlap region (see Liu et al., 2009), where (to within the accuracy and precision of our measurements) aerosols and clouds layer can have essentially identical optical properties, and thus cannot be distinguished based only on the CALIOP measurements. Within the overlap region, the best an algorithm can hope to achieve is to avoid biases by reporting roughly equal correct and incorrect classification rates. (Ideally, all 'overlap region' classifications will be flagged as "low-to-no confidence"; i.e., assigned very low CAD scores). Unfortunately, given the available layer attributes, FKM cannot help resolve issues in the overlap region. To better resolve cloud and aerosol layers, additional measurements and/or information are needed.

   Even the most sophisticated human observers cannot always agree on the correct partitioning of those layers that occupy the overlap regions (e.g., see Koren et al., 2007; Tackett and Di Girolamo, 2009; Varnai and Marshak, 2011; Balmes and Fu, 2018). These especially difficult cases include separating thin cirrus from lofted Asian dusts, separating evaporating water cloud filaments from the surrounding aerosols in the marine boundary layer, and separating fresh volcanic ash from cirrus. Given the measurements available on the CALIPSO platform, the classification of these targets is always subject to some uncertainty. So by using synthetic data to compare algorithm outputs versus "truth" we could perhaps choose an algorithm that best confirms our own

prejudices; but whether that algorithm was delivering the correct classifications in the really hard cases would still be an open question.

On to specific comments:

Abstract: As mentioned previously, the abstract gives minimal details about why you're undertaken this study, other than the rather vague "provide new insights" and validation of CAD. What are the new insights? Does it validate as expected?

We added more detailed results as "insight" in the abstract. We also clarified that in addition to validating the CALIPSO operational CAD algorithm (COCA), the comparison work helps establish the boundaries of classification correctness regarding the individual classification algorithm. The comparison works in general as good as expected with more than 94% of classifications agreeing between different algorithms.

Page 2, line 23: OK, so a difference between FKM and CAD is that the latter uses latitude as an input. I would think the arbitrary nature of the use of latitude is undesirable, so if FKM is successful than it's ability to perform without the use of latitude is very important and should be highlighted more.

The latitude information is used in operational V4 CAD algorithm to create the probability density function (PDF) to discriminate the cloud from the aerosol. For every five-degrees of latitude and 1km altitude, a look-up table of 3-dimentional probability cloud and aerosol is built according to the joint distributions of backscatter, depolarization and color ratio observation from lidar. This is because for different zones and altitudes, the sources and dynamics of cloud and aerosol are different. Applying a single, global scale look-up table would make it extremely difficult to identify local features. Latitude itself is not an intrinsic optical characteristic that can be universally used to separate cloud and aerosol. But because the probability for a particular class to be present is location dependent, latitude helps to shape the PDFs to provide better classifications at a local scales.

The reason we don't use latitude as an input is because we use FKM method, which is a centroid classification method, and geographic information about clouds and aerosols doesn't provide good separation criteria (they both occur at all latitudes everywhere around the planet) and hence will confuse the FKM classification. When we add latitude as an additional input, the resulting Wilk's lambda rises to 0.5, indicating that the classifications are no longer reliable.

Page 2: here you mention the difference between CAD V1 and CAD V4, but later on you validate against V3 and V4, and even mention V2 at some point. It seems overly complicated to compare against anything other than the latest version, but if you must you need to describe what is in each of the versions, the important differences, and why you need to validate against 3 & 4.

We added explanations about the differences between version 3 & 4 CAD algorithms. We also explain briefly why we have chosen to compare FKM to both CAD versions.

Note too that our manuscript is a part of an AMT special issue on CALIPSO version 4 algorithms and data products, and thus is a companion to the paper by Liu et al. (2018) that describes the V3-to-V4 updates to the CALIOP CAD algorithm. Our comparisons of FKM to both V3 and V4 CAD are meant to provide additional insights into the improved performance of the V4 CAD.

Page 4: OK, great, FKM doesn't use Latitude. Why not also skip altitude? Or try with and without? Altitude to me also seems like an arbitrary input that may bias your results, although perhaps somewhat more justifiable than latitude.

The relative occurrence frequencies of aerosols and clouds are quite different, depending on altitude (and, to a lesser extent, latitude), and thus altitude is likely to be a highly relevant characteristic for supervised learning approaches. How relevant it is for unsupervised learning is somewhat less obvious. (One might argue that mid-layer temperature would be a much better choice, as in some cases (e.g., $T > 0$ °C) temperature can be a determining factor in distinguishing between ice clouds and water clouds.)

However, the reviewer is right that altitude is also not an intrinsic optical property that can be used to discriminate between clouds and aerosols. Instead it is an atmospheric dynamics property: aerosols and water clouds are most often found in the boundary layer, while ice clouds can reach as high as the tropopause. So altitude can be crucial to discriminating between aerosol and ice clouds (though in some extreme cases their optical properties may be similar). For this reason we retain altitude in the FKM classifications. But, as the reviewer suggested, we have also experimented with omitting altitude for some particular cases in the paper (see section 4.4 in the revised manuscript). Throughout the paper we added more explanation saying why we choose these four parameters.

Page 5, last paragraph: You're using a random distribution of initial class memberships. How sensitive are you to that randomly selected distribution? Is there any difference in the results between one random seed and another? Also, is there any potential benefit in starting with class centers corresponding to preconceived notions of the class centers?

Actually, the code includes a loop for the selection of the random initiations so that the algorithm converges and gets the best fit. We now clarify this point in the text and in the flowchart in Figure 1. We did not check the impacts of initiation distributions on classification results in the paper. To response to the reviewer's suggestion, we did a quick check on this and found the initial scattering does not impact the results much (please see the figures below). We used the uniform and normal random distributions for a 2-classes clustering and the results are almost the same; the difference for one orbit is 0.00000816%. While we could also use preconceived class centers, I don't think this will change the class membership value, but it might change the speed that the algorithm converges.

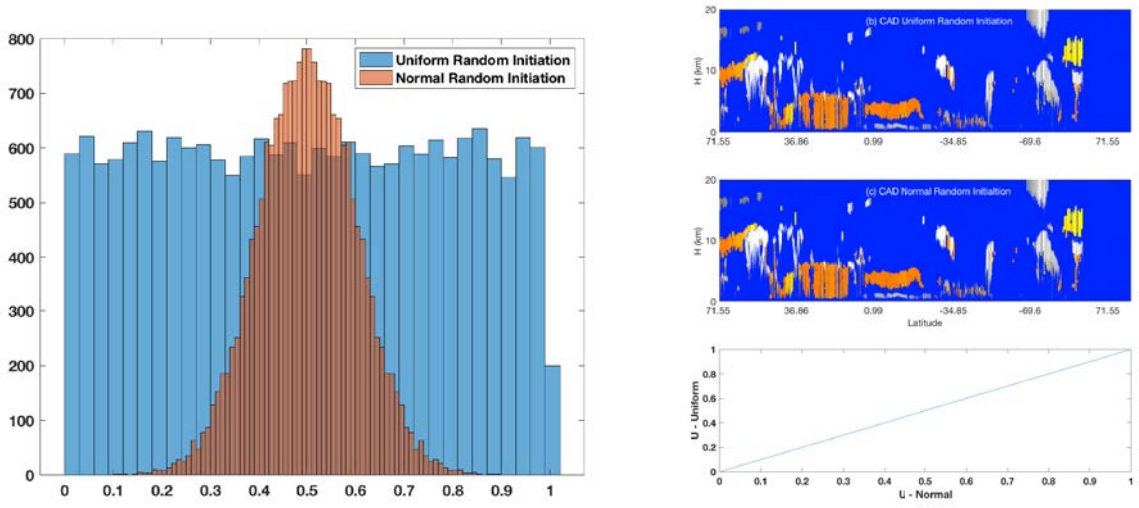


Figure 1, step 4: You have change in the norm of m as a metric to stop iteration (or max iteration number). Could other metrics also be used, such as change in c, or d? I've seen iterative methods that use multiple means of assessing when no further improvement can be provided as a means to reduce the number of time the max iteration number boundary is used.

We modified the flowchart and explanations about the last step (step 4) to determine our final membership due to our mistake in the previous manuscript. We actually used the objective function instead of membership which is not moving to determine when the iterations terminate. The code stops when algorithm converges and objective function change is smaller than a small threshold or the iteration count larger than a certain number. And we loop through multiple initiations and choose the smallest objective function minimum among all loop to get our final clusters. In real-world runs, the algorithm stops before reaching the max iteration number, which means the algorithm converged and objective function is stable. Note, the objective function is a function of membership, center and distance. That is why we say using only the center may slow down the speed. But I think technically we can use only the center, the distance or the membership. As the objective function takes joint account of all three, it would be our best choice.

Page 7, paragraph 1: could the FKM be used to improve the PDF used in CAD for the 'arbitrary' classification inputs (latitude and altitude)?

Perhaps…although to us it's not immediately clear how. In particular, the CAD altitude increments are chosen to partition the PDFs in a manner consistent with atmospheric dynamics. To a lesser extent, so too are the latitude increments. In fact, one of the primary reasons for increasing the latitude resolution from 10° in V3 to 5° in V4 was to achieve more reliable separation between ice clouds and dust in the northern hemisphere dust belt.

Page 7, second paragraph: "...sample is used to determine the optimal number of classes and fuzzy exponent required for classification. . ." How is this done? This is an important detail to skip.

We added more details about the selection of training data. We actually used one month of data (January 2008) to determine the optimal number of classes and fuzzy exponent required for classification.

Figure 2: how were these PDF's defined? Also, caption needs to spell out HOI and ROI

We added the definitions for PDFs in the text and the caption of Figure 2. We also add the definitions for HOI and ROI in the caption of Figure 2. The V4 CAD PDF is a five-dimension probability look-up table. Here the PDF is just the probability for one dimension, which is the occurrence frequency with sum normalized to unit.

Table 1: Based on the PDF's in Figure 2, it seems the filter criteria for AB is much tighter than what was selected for DR and CR. Why? Shouldn't they be similar?

Feature with AB larger than 0.2 are not majority of the cases and all of them are clouds. Aerosols with backscatter larger than 0.2 are relatively rare (e.g., only 3.5% of all tropospheric aerosol layers detected during 2012 at 5-km horizontal averaging had $<\beta'_{532}> > 0.2$ sr^{-1} km^{-1}). Setting the value as 0.2 or 2 won't change the results much. When we set the values larger than 0.2 to 2, it won't change the classification results, but it will speed up the calculations and make it easy to converge. We added some more explanation about this in the paper.

Page 9, paragraph 1: "Mahalanobis distance can be used for correlated variables. . ." This seems to imply that you expect AB, DR and CR to be correlated, but can you make it clearer if they are or not? If they are (and I suspect this is the case), this has implications for the uncertainty analysis later.

As you suspect, the lidar observables are indeed correlated. If we consider each of the three as sums (or means) of the measured backscatter signal over some altitude range, then

$$\langle\beta'_{532}\rangle = \frac{1}{N}\sum_{n=1}^{N}\beta'_{532,\parallel}(z_n)+\beta'_{532,\perp}(z_n), \quad \delta_v = \sum_{n=1}^{N}\beta'_{532,\perp}(z_n) \Big/ \sum_{n=1}^{N}\beta'_{532,\parallel}(z_n), \text{ and } \chi' = \langle\beta'_{1064}\rangle \Big/ \langle\beta'_{532}\rangle = \sum_{n=1}^{N}\beta'_{1064}(z_n) \Big/ \sum_{n=1}^{N}\beta'_{532}(z_n).$$

(The subscripts $\parallel$ and $\perp$ represent contributions from the 532 nm parallel and perpendicular channels, respectively.) In particular, the signals measured in the 532 nm parallel channel contribute to all three quantities. We now provide these details in (the new) section 3.4.

Figure 3: (a) x-axis title should have caps (b) what is NCE? And more generally how do we know from these figures that the ideal # of classes is 3 or 4 and corresponding fuzzy exponents 1.4 or 1.6. It is not clear what I should be looking for in these plots.

X-axis is the class number (k), and NCE is MPE instead. We modified them in figure and its caption. We choose the value based on that fuzzy exponents is "best value of ϕ for that class is at the first maximum of objective function curves", which in subfigure c, for 3 classes (red color) the pick value corresponds to ϕ =1.4 and for 4 classes (light green) ϕ = 1.4 or 1.6. We choose the class number based on the minimum values of both FPI and MPE and also real class number in the atmosphere (cloud and aerosol).

Page 11, second paragraph: mentioned before, but again: is it essential to compare against both V3 and V4? It seems to be unnecessarily complicated.

Not absolutely essential, perhaps, but certainly highly desirable from our point of view. See our previous response on this topic for a bit more detail.

Page 11, line 17 ". . .around 74S are misclassified. . ." confused by this statement since the lowest latitude on the figure is 71.44S

It actually extended to 71.55N and about 80S but the figure only shows 69.6S. We modified "around 74S" to "beyond 69.6S"

Page 13, line 15 "Note the value (1-CI)x100 for the 2-class FKM algorithm is equivalent to the CADFKM score" isn't it equivalent to the absolute value of the CADFKM score?

We modified "the CAD_{FKM} score" to "absolute value of the CAD_{FKM} score".

Page 14, line 6 "It is evident. . ." While I agree they do look this way, there are much more rigorous comparison 'statistic' out there than eyeballing a figure.

We added statistics to support our "evident" claim.

Figure 7: it appears there is a big difference at high latitudes between FKM and CAD for clouds. Considering that CAD uses latitude, isn't this a very significant difference that should be highlighted and discussed further?

Further analysis of CAD in high latitude is out of the scope of work at the moment. In the future we can apply FKM for different altitude, latitude and season so as to improve the classification at local level. For different latitudes, the atmosphere dynamics are different, also the clouds and aerosol source and their intrinsic properties are quite different as well. To solve the problem, we can't simplify the problem by just adding/removing latitude with equal weighting. In this case sampled here, maybe the class number is not sufficient for the classification, so we may improve by adding classes number due to the large differences between stratospheric cloud and aerosol. This problem should be addressed in the future. Right now, the studies mainly discuss the comparisons of the classifications in the troposphere. We just want to highlight this information in the paper at the moment instead of trying to optimize FKM performance to resolve various problems. The paper already has a lot of information, and needs to stop somewhere so that we do not lose our focus (i.e., assessing the performance of the CALIOP operational CAD algorithm).

Table 2: What should I be looking for in the C and A columns and rows? This would be more meaningful if I knew what the actual expectations for percentage C and A should be.

The clouds global distribution is between 50%~70% according to Stubenrauch et al. (2013). For all algorithms or versions discussed in the paper, the cloud coverages are well within ranges that are considered acceptable. We added this reference in the paper text to support the comparison work here.

Figures 4, 8, 9, 10, 11 would be easier to understand if the colorbar/label at the bottom indicated which side was 'aerosol like' and which was 'cloud like'.

We add "cloud" and "aerosol" below the color bars to indicate cloud-like and aerosol-like colors. Since we still want to show the confidence of the classification, we added the following text just before

Figure 8: "CAD classifications are color-coded as follows: regions where no features were detected are shown in pure blue; fill values 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 gray scale, with brighter and whiter hues indicating higher classification confidence."

Page 19, lines 12-13. This is an important point that must be emphasized! Use of altitude leads to mis-classification!

Yes, in some cases, using altitude will lead more misclassifications (e.g., in regimes having roughly equal numbers of samples of high altitude aerosol and ice clouds that also have highly similar optical properties) but for other cases (i.e., for a majority of cases, when separating ice clouds and low altitude aerosols) it will improve the classification. This is now discussed briefly in the last section. Altitude adds more confusion in separating water clouds from aerosols (e.g., when water cloud backscatter is small in optically thin water clouds), and in separating PSCs from stratospheric aerosols. Because these layers are at similar altitudes, altitude does not provide useful distinguishing information. But on the other hand, the altitude information can improve the classification between ice clouds and aerosols even if their depolarization and backscatter are similar. Ice clouds are high-altitude features, whereas aerosols are most often much lower. So it is hard to categorically say that the use of "altitude leads to misclassification". From Table 4, we also know that the combination of altitude and color ratio provides the best FKM classification compared to the V4 operational classification. It depends on the cases sampled and studied. While backscatter intensity can also produce confusion in the separation of thin clouds and dense aerosols, this occurs in much fewer cases. In majority of the cases, altitude helps improve the separation. We don't want to make an absolute conclusion just because of this case.

Page 21, line 6: this is the first time we hear that a reason for the use of V4 is improved calibration coefficients. Version differences for CAD should be described in more detail in an earlier section.

We added some brief explanations of the differences between V3 and V4 in the introduction. Detailed explanations of the differences between V3 and V4 are given in Liu et al. (2018), which we expect to be published part of this same AMT special issue. Our focus in this manuscript is on FKM and the comparison of the operational products with FKM products.

Figure 10 caption: spell out PSC, STS and NAT acronyms

We deleted the subsection discussing PSCs because the primary focus of our paper is on comparisons in troposphere.

Page 23, line 14: this is a confusing statement – you're using FKM to rebuild PDFs?

The referee is 100% correct: as originally written, this is indeed a confusing statement. To clarify, we have replaced the original text with the following revision.

Original

> With the FKM method, it is easy to add or remove one or multiple observational dimensions and re-cluster without re-building new PDFs.

Revised

With the FKM method, it is relatively easy (though perhaps time-consuming) to add or remove one or multiple observational dimensions (i.e., inputs dimensions) and the reinitiate the training/learning algorithm. (This highly desirable flexibility is, unfortunately, wholly absent in the strictly supervised learning regime incorporated into COCA.)

Page 23, line 20: under what circumstances does classification degrade with the addition of additional dimensions?

According to my understanding, it really depends on the case. This is also explained in the last section for the EOF analysis. If the dimension does not contribute information that helps segregate the data into distinct classes, it may instead degrade the accuracy of the classification. As an entirely speculative example, suppose we were to add UTC time-of-day as an additional input dimension. Since CALIPSO has a 16-day orbit repeat cycle, this sort of "information" may introduce subtle (or even not so subtle) artifacts into the derived clusters. To provide further clarification on this issue, we have added a pointer to additional discussion later in the manuscript (i.e., "details provided in Sect. 4.3").

Figure 11: What is HH? I'm assuming altitude, but you use H elsewhere for that. It's probably best to spell out AB, DR, CR, HH in caption.

We spelled out AB, DR, CR and H in the caption and change the "AB, DR, CR, and HH" in the title to "$<\beta'_{532}>$, δ_v, χ', and z_{mid}" to keep all symbols in this paper same as those in the CAD paper published by Liu et al. 2018.

Page 25: I really don't understand why you are avoiding high latitude and altitude regions. This is where FKM could presumably help, and differences with CAD might indicate problems with the latter's use of altitude and latitude as dimensions. Or, it might indicate that FKM without altitude and latitude still can't resolve well, in which case those dimensions are needed with CAD. This seems like a in important issue you're sidestepping.

At high altitudes or latitudes, the intrinsic properties of clouds and aerosols are often much different from those in the other areas of the globe. Consequently, 2 or 3 classes may be not enough for the classification at a global scale. To improve the performance, we most likely need to apply FKM at local scales. And while that has not been done in this paper, it is worth further study in the future. This paper is not meant to be an exhaustive analysis, but should instead be seen as an initial step in an on-going CAD validation study. Future, more focused investigations of CAD performance in polar regions and the stratosphere are highly desirable next steps.

Figure 12: How can you expect anybody to understand this figure? After a bit of effort, I think I understand what you're trying to show, but even if I am correct there's no way for me to differentiate the 15+ different colors. What should this figure look like in a perfect case? I think this figure and the corresponding text are an example of something that should be cut so more focus can be given to other sections.

We deleted the figure. Instead, we used a table to summarize the numbers so as they are more easily understood by the reviewers and other readers. These numbers provide information all of the

information originally presented in the figure (e.g., how much improvement is made if we changed the inputs dimension from 3 parameters to 4 parameters, etc.).

Equation 16: It would be nice if you said a sentence about what this matrix should look like (i.e square pxp matrix that is I in a perfect case)

We added more the dimension information in the paper text i.e. "(k x p matrices)" in the text.

Equation 18: What kind of norm is this?

Equation 18 is not a norm, but instead shows the ratio of the determinant of W and the determinant of W + B. To clarify this, we have replaced the |W| notation with det(W). Determinates represent the volume of matrices in multiple dimensions while norms quantify distance in multiple dimensions.

Table 3. I'm having difficulty interpreting this. We want low Wilks' lambda for best classification, right? So lowest values are for backscatter alone for 2 class, and backscatter and depol for 3 class. So, does higher lambda when adding other parameters mean classification become worse? Or is it that wilks lambda can't be compared when the dimensionality is different? This is an example of an analysis that seems lacking in its description of what you are looking for, and what the results mean.

The best classification needs to consider many factors. The Wilks' lambda is only one indicator to help assess whether the clusters are sufficiently distinct or not. If clusters have no boundary between each other, although the classification accuracies are high, the classification results can't be trusted either. The classification results can be trusted when the accuracy is high as well as the Wilks' lambda is small, namely clusters are distinct to each other. I don't think that the dimension number itself will cause the increase of Wilk's lambda, it is how well a certain dimension bring in useful information for the distinction of the classification that matters. For example, the dimension of altitude adds more ambiguity to the classification confidence for the 2-classes classification because clouds are found at both high and low altitudes while aerosol are mostly found at low altitudes. For majority of the cases the backscatter intensity and depolarization ratio won't contribute ambiguity for the 2-classes classification and the backscatter intensity and color ratio won't contribute ambiguity for the 3-classes classification. We have to jointly looking at how well the classification does according to all the subsections in the sections. Generally speaking, smaller values indicate higher confidence in the results. But once the dimension brings in fuzzy factor (to make the boundary overlap a lot), although the accuracy increases, the classification Wilks' lambda will decreases. We added more explanation in the paper to clarify this.

Section 4.3: why not just do PCA on the input parameters?

We wanted to see what makes the classification distinct, so we have to use output instead of input to do PCA. We use Wilk's lambda to do PCA to see how well the classification results are. If we use inputs to do PCA, it may indicate how independent the inputs are which will answers the question on page 9 paragraph 1 from the reviewer. We reorganized the phrase to make it clear.

Figure 13: I'm confused by the distribution here. Are we to understand that aerosols are a class completely (or mostly) surrounded by other classes? Also, why these axis ranges, at least make PCA (2) from -4 to 4 so we can see what is going on.

It doesn't means that aerosol is completely surrounded by other classes. It just means aerosols are completely separated classes, as you can see the more condensed samples (darker colors, or centers in red crosses) of each class are distinct from each other. For example, thin clouds and dense aerosols are more similar to each other, so the aerosol cluster will have overlapped zone with clouds but the overlapped samples are less (lighter colors). We added more details about the colorbar in the paper to explain this. To let reviewer to clearly see the relationship between PCA2 and PCA1, we have reoriented the axis. This is because PCA1 contribution to the classification is more significant compared to PCA2 so that C1-C2 line is approximately in diagonal when PCA1 and PCA2 contribution is equal. We added more details in the paper.

Page 30, line 11: so, you're using one second of observations for the error propagation? I understand the computational limitations but this seems exceedingly limited.

Yes, the original text was not clear on that point. What we claimed was that "the observations from 6 September 2008 at ~01:35:29 UTC are used", and certainly that implies that we used only one second of data. What we used, however, was a full granule of data. Our revised sentence now says "the observations from ***a nighttime granule acquired*** 6 September 2008 ***beginning*** at ~01:35:29 UTC are used"

Section 4.4 Aren't your dimensions correlated to some extent, such that you should expect uncertainties to be related (correlated) as well?

Yes, the inputs are correlated, so the covariance terms in the uncertainties could either magnify or mask our current uncertainty estimates. The paper is not designed to check those.

Equation 19: What kind of norm?

Equation 19 is the L1 norm. We now say so explicitly in the revised manuscript.

Figure 14: why not do this with actual CALIOP uncertainties, since these have been assessed? Also, why not use all three uncertainties simultaneously, like the real world?

Figure 14 used the actual CALIOP uncertainties for each CALIOP observations. We also added, as the reviewer suggested, the results from the combination of the three uncertainties occurring simultaneously, as would be seen in real world.

Figure 15: why is the 2-class case the only one investigated this way?

Because the paper mainly discusses the CAD, which is a two class partitioning of the identified layers.

Page 33, like 28: "While the two algorithms use largely identical inputs..." I heartily disagree with this statement. CAD uses altitude and latitude, which your analysis has shown to be important (even if you don't emphasize as much as I would like).

The revised text now says "While the two algorithms both rely on the same underlying lidar measurements…", which is entirely accurate.

References

Balmes, K. A., and Fu, Q.: An Investigation of Optically Very Thin Ice Clouds from Ground-Based ARM Raman Lidars, Atmosphere, 9, 445; doi:10.3390/atmos9110445, 2018.

Brakhasi, F., Matkan, A., Hajeb, M., and Khoshelham, K.: Atmospheric scene classification using CALIPSO spaceborne lidar measurements in the Middle East and North Africa (MENA), and India, International Journal of Applied Earth Observation and Geoinformation, 73, 721-735, https://doi.org/10.1016/j.jag.2018.07.017, 2018.

Koren, I., Remer, L. A., Kaufman, Y. J., Rudich, Y., and Martins, J. V.: On the twilight zone between clouds and aerosols, Geophys. Res. Lett., 34, L08805, doi:10.1029/2007GL029253, 2007.

Liu, Z., Vaughan, M., Winker, D., Kittaka, C., Getzewich, B., Kuehn, R., Omar, A., Powell, K., Trepte, C., and Hostetler, C.: The CALIPSO lidar cloud and aerosol discrimination: Version 2 algorithm and initial assessment of performance, J. Atmos. Oceanic Technol., 26, 1198-1213, https://doi.org/10.1175/2009JTECHA1229.1, 2009.

Liu, Z., Kar, J., Zeng, S., Tackett, J., Vaughan, M., Avery, M., Pelon, J., Getzewich, B., Lee, K.-P., Magill, B., Omar, A., Lucker, P., Trepte, C., and Winker, D.: Discriminating Between Clouds and Aerosols in the CALIOP Version 4.1 Data Products, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-190, in review, 2018.

Tackett, J. L., and Di Girolamo, L.: Enhanced aerosol backscatter adjacent to tropical trade wind clouds revealed by satellite-based lidar, Geophys. Res. Lett., 36, L14804, https://doi.org/10.1029/2009GL039264, 2009.

Varnai, T. and Marshak, A.: Global CALIPSO Observations of Aerosol Changes Near Clouds, IEEE Geosci. Remote Sens. Lett., 8, 19-23, https://doi.org/10.1109/LGRS.2010.2049982, 2011.

Interactive comment on "Application of High-Dimensional Fuzzy K-means Cluster Analysis to CALIOP/CALIPSO Version 4.1 Cloud-Aerosol Discrimination" *by* Shan Zeng et al.

Anonymous Referee #2

Received and published: 28 September 2018

The manuscript describes a methodology to discriminate between aerosol and cloud layers from CALIOP/CALIPSO lidar Level 2 data based on the high dimensional Fuzzy K-Means Cluster Analysis. The argument for sure is a good fit for the journal but some parts are not clear, probably suffering from hasty writing and need improvements before final publication. Moreover, other tests should be performed to improve scientific significance and clarity. I am however confident that the authors will brilliantly address all the issues I raised.

Thanks for reviewing the paper and giving valuable feedback. It is very hard to validate the operational algorithm at global scale, because we know of no existing global in-situ data set that could be used for the task. A comparison between different classification schemes used by active and passive sensor has been done in previous work (Stubenrauch et al., 2013). However, as active sensors profile the full vertical extent of the atmosphere, it remains quite difficult to compare classification results with passive sensors that, at best, only measure the properties of a single layer. (More often, properties of multiple layers are convolved into a single set of measurements, and thus tasks such as separately classifying cirrus clouds and boundary layer aerosols within the same pixel are extremely challenging retrievals for passive sensors.) Furthermore, comparisons between different algorithms have not yet been performed. Similar to the comparison between passive and active sensors, it's hard to determine how accurate the algorithms are (see our previous comments about the use of synthetic data), but by combining data from multiple sensors we can estimate upper and lower boundaries for cloud and aerosol distributions over the globe, and these values give a distribution range to guide modelers. Similarly, the comparison between supervised and unsupervised algorithms can also give upper and lower boundaries for precision to guide modelers, instrument developers, and data processors. To address the referee's concerns, we add detailed statements in the introduction to clarify these points. As we say in the conclusions of the original draft, the purpose of this study is "to validate the performance of the cloud-aerosol discrimination (CAD) algorithm used in the standard processing", and we are not suggesting FKM as a replacement for COCA. To this end we also added more introduction about the importance of discriminating between clouds and aerosols, and described the benefits for the study for different user communities.

Major Comments:

The FKM clustering methodology is well described and totally makes sense. But, as stated in the introduction, the FKM method is used to validate the result of V4 CAD algorithm and to better understand the classification, identifying the crucial parameters. It looks like that all the produced efforts have a very low return on investment. The V4 CAD is not validated vs. a reference dataset, i.e. using a synthetic lidar data where all the aerosol and cloud properties are well known and controlled, but with respect to another methodology that have comparable uncertainties.

While the use of synthetic data as an evaluation tool is generally an excellent and highly effective strategy, in the case of discriminating clouds from aerosols it's also especially hard to implement in a useful way. For ~90% of the cases, cloud and aerosol properties are very well separated and reliable classifications can be made using a single wavelength elastic backscatter lidar (e.g., CATS; see https://cats.gsfc.nasa.gov/media/docs/CATS_QS_L2O_Layer_3.00.pdf). In these cases, unambiguous synthetic data can be used to weed out those algorithms that are obviously deficient. But the remaining cases fall into the cloud-aerosol overlap region (see Liu et al., 2009), and for these layers even the most sophisticated human observers cannot always agree on the correct partitioning (e.g., see Koren et al., 2007; Tackett and Di Girolamo, 2009; Varnai and Marshak, 2011; Balmes and Fu, 2018). These especially difficult cases include separating thin cirrus from lofted Asian dusts, separating evaporating water cloud filaments from the surrounding aerosols in the marine boundary layer, and separating fresh volcanic ash from cirrus. Given the measurements available on the CALIPSO platform, the classification of these targets is always subject to some uncertainty. So, yes, we certainly could create "synthetic lidar data where all the aerosol and cloud properties are well known and controlled" and compare the classifications obtained from the CALIPSO operational CAD algorithm (COCA) and the FKM algorithm. And by using this synthetic data to compare algorithm outputs versus "truth" we could perhaps choose an algorithm that best confirms our own prejudices; but whether that algorithm was actually delivering the correct classifications in the really hard cases would still be an open question.

Moreover, It is completely missing an analysis on who is really using those data, i.e. climatologists, modelers. . ., and why it is critical to discriminate (defining a level of precision) between aerosols and clouds (and their subtypes). For example, how much is it the actual precision of the current operational V4 CAD algorithm in classifying the aerosol and cloud layers ? The final users are ok with this accuracy? Which benefits will be obtained reducing the misclassification? How the FKM will be used or implemented to reduce the V4 CAD misclassification?

While this would certainly be interesting information, this kind of detailed analysis lies well beyond the scope of this paper. (Simply counting up the number of different CALIPSO data user communities that make use of the CAD scores we provide would likely lead to some fascinating (and perhaps surprising!) insights.) In this paper, our goal is limited to providing a performance assessment of the current CALIPSO operational CAD algorithm.

In the manuscript is only marginally discussed why January 2008 measurement are a representative data sample. How the results are impacted changing the analyzed dataset?

As one-month data is enough for the purpose of the study, we just randomly choose one month. For different dataset, the class number and the fuzzy exponent may be different, but classification results on cloud and aerosol should not be too different in theory. In reality, for different season, different features occur which may slightly impact the sample of classes and thus the results. The paper just focuses on the first step of comparison and didn't go further. We mentioned this in the data preparation and added a summary of future work at the end of conclusion.

The number of classes is predefined (2 or 3) after analyzing Figure 3. However, in operational contexts, some data subsets might belong only to two classes. FKM still will fill with observation the class that should be empty. Is there a reason why the authors used the FKM cluster analysis instead of some self-selecting class methods, i.e. MeanShift clustering (Cheng, Yizong. "Mean shift, mode seeking, and clustering." IEEE transactions on pattern analysis and machine intelligence 17.8 (1995): 790-799) or classification algorithms as AdaBoost (Hu, Weiming, Wei Hu, and Steve Maybank.

"AdaBoost-based algorithm for network intrusion detection." IEEE Transactions on Sys- tems, Man, and Cybernetics, Part B (Cybernetics) 38.2 (2008): 577-583) ?

I think both Meanshift and Adaboost are very good algorithms to do clustering too. There are so many clustering methods (more than 100 maybe), supervised or unsupervised, connectivity-based, centroid based, density based and distribution clustering, we only try the Fuzzy K-means out, which is one of unsupervised centroid method that produces a membership which (between 0 and 1) is represent the probability of belonging to one class and is comparable to the official CAD scores (between -100 to 100, probability belong to one and the other). And also the shape of multi-dimensional observations of cloud and aerosol are suitable for centroid based algorithms. Last, the FKM unsupervised approach is quite different from the highly supervised method used to train the operational algorithm, is what we need for the comparison and the objective of the study.

Density based algorithms such as Meanshift expect some kind of density drop to detect cluster borders. Mean-shift is usually slower than k-Means. Besides that, the applicability of the mean-shift algorithm to multidimensional data is hindered by the unsmooth behavior of the kernel density estimate, which results in over-fragmentation of cluster tails (Achert et al. 2006). Clouds have two centers (ice and water) and aerosols may also have several sub-centers (e.g., dust and biomass burning), so a density based algorithm may not suitable for this classification in my opinion. Also, according to Kaur and Chawla (2015), FCM has higher accuracy compared to the Meanshift. AdaBoost is a machine learning method, and more complicated to understand. While using it may resolve the problem for FKM weighting problems in some future study, at the moment we want an easier understand method that is distinctly different from the COCA method investigate different algorithm inputs on the classifications. In the future we will consider to doing some machine learning classifications, but may not choose AdaBoost.

The random initialization of the centroids is a well-known problem as the initial centroid selection not only influences the efficiency of the algorithm, but also the number of relative iterations (and consequently the needed time machine). Some optimal centroid selection techniques can be found in Nazeer, K.A. Sebastian, M.P Clustering biological data using enhanced k-means algorithm". In: Electronic Engineering and Computing Technology, Springer, 2010, pp. 433–442 (chapter 37)

The flowchart is wrong in previous version of manuscript. We have a loop to choose the best initiation and outcome results in FKM algorithm. With the loop to choose the best initiation, the larger the number of loops, the better the resulting clusters will be, but this is not time efficient. In application to real data, we have not yet found that using a larger number of loops will consistently improve the classification accuracies for the CALIPSO level 2 observations.

Many thanks to the reviewer for introducing us to an efficient way to save the relative iteration number and time.

Specific comments:

Line 27 Pag. 1 Please add also "geometrical properties"

We added it.

Line 15 Pag. 5 How the random initialization influence the final result? I don't recall any section where this issue is discussed. Are the results consistent with the random initialization?

We misrepresented our algorithm, and so we modified our flowchart in Figure 1. As we do a loop to choose the best random initialization, outcome results do not change due to initiation as long as the iteration number and the loop number for selecting initiation are big enough.

Line 16 Pag. 5 the authors mean Equations 2, 3 and 4?

We corrected them.

Figure 1: Third step it should be Eq. 6 and 7

We corrected them.

Line 2 Pag. 7: I am not sure that latitude is not useful to discriminate, as clouds at 16 km at polar latitudes may rise a flag, as cirrus clouds below 9km in the equatorial and tropical regions

The region (i.e. latitude) and season information are of course useful auxiliary information because they can indicate the sources of particles and the dynamics of the atmosphere. The others are directly measured optical information of the particles due to their scattering nature. In the future, we can train and apply the FKM method at local scales, which could be a way to improve the current classifications.

Figure 3: labels are difficult to read. The picture in the middle shows "NCE" that is not previously defined.

We selected the bold font to the labels so as to see the label easier and changed the "NCE" to "MPE".

Line 14 Pag 11: please rephrase "water clouds. For these water clouds".

We rephrased it.

Figure 4: it is very hard to see the zone of interest (smoke and cloud). Maybe reduce the vertical scale from 0 to 20 km?

We modified it to 20km.

Line 15 Pag 17 please read "We saw" instead of "We see"

We corrected it.

Paragraphs 3.4 a, 3.4 b and 3.4 c. How the authors assume that the layer are pure dust, smoke and ash respectively? Is there any other ancillary measurement that shows without any doubt the aerosol layer composition?

This comment highlights one of the major difficulties in validating a global data set acquired by a first-of-its-kind active sensor: coincident measurements of interesting phenomena are extremely difficult to come by! For these events, we tracked these plumes by eye according to the event's location, time period and our experience in evaluating spatial distributions and layer optical features (depolarization, color ratio and backscatter). This is very accurate though.

Section 4. Figure 13 is not very intuitive and it is difficult to get meaningful information from it . It might be interesting to replace it (or add) the Screen Plot and the loading factors as barplot as showed in https://doi.org/10.1175/JTECH-D-15-0085.1.

The figure includes a lot of information compared to the barplot, but we did not explain it well. We have now added more explanation about the figures and added the color bar.

Line 4 Pag. 34: Even if the FKM Cluster Analysis closely replicate the CAD V4 operational algorithm, it is not validate it (see main comment section)

We changed the "validation" to "comparison". We explained more in the paper that the comparison between algorithms can set up boundaries for the uncertainness due to different algorithms

Line 18 Pag. 35. FKM it is a time consuming algorithm because setting up random centroids can slow down the convergence process and in some cases can produce as result sub-optimal centroids virtual centroids (i.e. not corresponding to any observational measurement). See Main Comments section.

Yes, we added more details to the related domain to clarify the reason for FKM "time consuming". We modify the algorithm description in the paper and in Figure 1.

References

Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 3918: 119–128. doi:10.1007/11731139_16. ISBN 978-3-540-33206-0.

Balmes, K. A., and Fu, Q.: An Investigation of Optically Very Thin Ice Clouds from Ground-Based ARM Raman Lidars, Atmosphere, 9, 445; doi:10.3390/atmos9110445, 2018.

Brakhasi, F., Matkan, A., Hajeb, M., and Khoshelham, K.: Atmospheric scene classification using CALIPSO spaceborne lidar measurements in the Middle East and North Africa (MENA), and India, International Journal of Applied Earth Observation and Geoinformation, 73, 721-735, https://doi.org/10.1016/j.jag.2018.07.017, 2018.

Kaur, S. and Chawla S. (2015): Evaluation of Performance of Fuzzy C Means and Mean Shift based Segmentation for Multi-Spectral Images, International Journal of Computer Applications 120, 25-28.

Koren, I., Remer, L. A., Kaufman, Y. J., Rudich, Y., and Martins, J. V.: On the twilight zone between clouds and aerosols, Geophys. Res. Lett., 34, L08805, doi:10.1029/2007GL029253, 2007.

Liu, Z., Vaughan, M., Winker, D., Kittaka, C., Getzewich, B., Kuehn, R., Omar, A., Powell, K., Trepte, C., and Hostetler, C.: The CALIPSO lidar cloud and aerosol discrimination: Version 2 algorithm and initial assessment of performance, J. Atmos. Oceanic Technol., 26, 1198-1213, https://doi.org/10.1175/2009JTECHA1229.1, 2009.

Liu, Z., Kar, J., Zeng, S., Tackett, J., Vaughan, M., Avery, M., Pelon, J., Getzewich, B., Lee, K.-P., Magill, B., Omar, A., Lucker, P., Trepte, C., and Winker, D.: Discriminating Between Clouds and Aerosols in

the CALIOP Version 4.1 Data Products, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-190, in review, 2018.

Tackett, J. L., and Di Girolamo, L.: Enhanced aerosol backscatter adjacent to tropical trade wind clouds revealed by satellite-based lidar, Geophys. Res. Lett., 36, L14804, https://doi.org/10.1029/2009GL039264, 2009.

Varnai, T. and Marshak, A.: Global CALIPSO Observations of Aerosol Changes Near Clouds, IEEE Geosci. Remote Sens. Lett., 8, 19-23, https://doi.org/10.1109/LGRS.2010.2049982, 2011.