文章编号：1001-9014(XXXX)XX-0001-12

DOI：10. 11972/j. issn. 1001-9014. XXXX. XX. 001

基于CALIOP数据的气溶胶垂直分布特征聚类分析

王宇轩[1,2,3]，孙晓兵[1*]，提汝芳[1]，黄红莲[1]，刘 晓[1]，余海啸[1]

（1. 中国科学院合肥物质科学研究院 光学定量遥感安徽省重点实验室，安徽 合肥 230031；
2. 中国科学技术大学，安徽 合肥 230026；
3. 河南理工大学，河南 焦作 454002）

摘要：大气气溶胶的垂直分布具有高度复杂特征和时空变异性，是提高卫星遥感气溶胶反演效果的关键影响因素。研究基于2010年至2020年的CALIOP（The Cloud-Aerosol Lidar with Orthogonal Polarization）L3气溶胶剖面数据，采用无监督聚类方法对气溶胶的垂直分布特性进行了系统性研究。通过多个评估指标比较GMM（Gaussian Mixture Model）、K-means、谱聚类三种聚类算法的聚类效果。基于消光系数的垂直分布特征使用GMM聚类方法将气溶胶剖面划分为五种具有代表性的类型：低污染组合型、高污染组合型、指数衰减型、低污染均匀型和高污染振荡型。进一步分析了这些剖面在不同季节以及在青藏高原、京津冀、长三角三个典型地区的时空分布特征。研究结果表明，通过GMM聚类分析得到的气溶胶剖面呈现出显著的季节性和地域性差异。

关 键 词：气溶胶；气溶胶垂直分布；聚类分析；CALIPSO

中图分类号：P413　　**文献标识码**：A

Clustering analysis of aerosol vertical distribution characteristics based on CALIOP data

WANG Yu-Xuan[1,2,3]，SUN Xiao-Bing[1*]，TI Ru-Fang[1]，HUANG Hong-Lian[1]，YU Hai-Xiao[1]

（1. Anhui Province Key Laboratory of Optical Quantitative Remote Sensing，Hefei Institutes of Physical Science，Chinese Academy of Sciences，Hefei 230031，China；
2. University of Science and Technology of China，Hefei 230026，China；
3. Henan Polytechnic University，Jiaozuo 454002，China）

Abstract：The vertical distribution of aerosols plays a critical role in improving the accuracy of aerosol retrieval in satellite remote sensing due to its complexity and spatiotemporal variability. This study investigated the vertical characteristics of aerosols using unsupervised clustering methods，based on CALIOP（Cloud-Aerosol Lidar with Orthogonal Polarization）Level 3 aerosol profile data from 2010 to 2020. Three clustering algorithms—Gaussian Mixture Model（GMM），K-means，and spectral clustering—were evaluated using multiple performance metrics. The profiles of extinction coefficients were clustered into five representative types using the GMM algorithm：low-pollution composite type，high-pollution composite type，exponential decay type，low-pollution uniform type，and high-pollution oscillatory type. The seasonal and regional distributions of these profile types were further analyzed over the Tibetan Plateau，the Beijing-Tianjin-Hebei region，and the Yangtze River Delta. The results show that aerosol vertical profiles exhibit distinct seasonal and regional patterns. These findings provide a basis for improving aerosol profile parameterization and retrieval accuracy in remote sensing applications.

Key words：aerosols，aerosol vertical distribution，cluster analysis，CALIPSO

收稿日期：2025-01-03，**修回日期**：2025-02-26　　**Received date**：2025-01-03，**Revised date**：2025-02-26

基金项目：航天科技创新应用研究项目（E23Y0H555S1）、航空科技创新应用研究项目（62502510201）、中国科学院重点实验室基金项目（E33Y0HB42P1）

Foundation items：Supported by the Aerospace Science and Technology Innovation Application Research Project（E23Y0H555S1），the Aviation Science and Technology Innovation Application Research Project（62502510201），the Chinese Academy of Sciences key Laboratory Fund Program（E33Y0HB42P1）

作者简介（Biography）：王宇轩（1996—），男，河南开封人，在读博士，主要研究领域为气溶胶遥感. E-mail：zy070030@mail. ustc. edu. cn

[*] **通讯作者**（Corresponding author）：E-mail：xbsun@aiofm. ac. cn

引言

气溶胶是指悬浮在大气中的微小固体或液体颗粒，其粒径通常在0.01~10 μm之间。气溶胶的来源包括自然过程（如火山喷发、沙尘暴等）以及人类活动（如工业生产、燃料燃烧等）[1]。近年来，随着工业化和城市化进程的加速，人类活动所产生的气溶胶排放量显著增加。气溶胶通过直接和间接辐射效应改变地球的辐射收支，在加剧空气污染的同时，对全球气候系统也产生了深远而复杂的影响[2]。此外，粒径尺寸小于10 μm的气溶胶颗粒可以深入人体呼吸系统，导致一系列健康问题，包括呼吸道疾病、心血管疾病以及肺部感染等[3-4]。气溶胶因其组分多样，导致粒子的微物理与光学特性复杂，加之其高度的时空变异性，使得其对环境和健康影响的定量化评估仍然存在挑战。

气溶胶测量中一个较为关键的参数是气溶胶的垂直分布[5]。气溶胶的垂直分布不仅对大气边界层结构的稳定性具有重要影响，还直接影响云与气溶胶的相互作用，改变降水模式，并显著影响空气质量[6-7]。此外，目前气候模式对气溶胶垂直剖面的模拟结果可能与真实情况相差一个数量级，这使得估算气溶胶辐射效应时产生显著误差，进而影响反演算法开发时先验知识的准确性[8-9]。气溶胶的垂直分布在时间和空间上均表现出高度变化性，其分布特征受到风速、降水、大气条件以及气溶胶粒子微物理与光学特性等多种因素的共同影响[10]。例如，位于近地表边界层的气溶胶粒子可以被迅速输送到更高海拔的区域，这一动态过程进一步加剧了气溶胶分布的时空复杂性。上述因素共同导致气溶胶遥感中的显著不确定性[11]。

近年来，诸多学者利用地基或星载激光雷达开展了一系列区域性研究，以更好地理解大气中气溶胶的垂直分布结构及其光学性质[12-13]。固定式激光雷达能够提供高频次、精确的气溶胶垂直分布观测：地基激光雷达则适用于单个站点的高精度探测；星载激光雷达则可以实现对气溶胶的全球时空连续覆盖观测。其中，2006年发射的CALIPSO（Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations）极轨卫星是一个极为重要的数据源，其搭载的CALIOP仪器能够提供三维云和气溶胶产品[14]。Yu等人利用一年CALIOP数据研究了气溶胶垂直分布的季节变化，结果显示CALIPSO对气溶胶光学厚度（Aerosol Optical Depth，AOD）地理分布和季节性变化的观测与GOCART（Georgia Tech/Goddard Global Ozone Chemistry Aerosol Radiation and Transport）模型模拟、MODIS（Moderate Resolution Imaging Spectroradiometer）反演结果具有良好一致性[15]。Winker等人基于多年CALIPSO L3气溶胶产品，构建了全球对流层气溶胶的三维分布模型。研究结果表明，由于气溶胶排放源强度和输送机制的显著差异，其垂直分布具有明显的季节性变化[16]。

此外，随着机器学习技术的不断进步，聚类分析因其不依赖先验类别、适用于处理复杂和高维遥感数据的特性，近年来在大气科学和遥感研究中获得广泛关注和应用。Fan等人利用AERONET（Aerosol Robotic Network）站点的数据，通过聚类分析得到了8种典型气溶胶模型，并研究了这些模型在典型地区的季节变化特征[17]。Wang等人结合CALIOP数据和PBLH（Planetary Boundary Layer Height）数据，采用模糊K均值方法对CALIOP的10年气溶胶剖面数据进行了空间聚类分析，根据目视效果将气溶胶分为污染簇、中等簇和清洁簇，并对各簇特性进行了详细分析[18]。总的来说，聚类分析作为一种无监督学习方法，可以根据数据的相似性或差异性，将观测数据自动分组，从而挖掘气溶胶特性在空间、时间或高度上的分布规律[19]。

本研究基于10年的CALIPSO气溶胶剖面数据，采用聚类分析方法对中国大陆地区的气溶胶剖面进行了分类研究。聚类方法不仅有助于揭示气溶胶剖面在不同地区的空间分布特征，还可用于深入分析其垂直结构及潜在驱动因素。为了更准确地揭示气溶胶的垂直分布模式，分别对原始气溶胶剖面数据和归一化特征提取处理后的数据进行了分析。本文的结构安排如下：第一章详细介绍了使用的数据来源，聚类分析的具体方法以及数据的预处理步骤；第二章展示了聚类分析的结果，重点讨论了气溶胶剖面的季节变化规律和空间分布特征；第三章总结了研究结论，并对未来工作进行了展望。

1 数据与方法

1.1 研究数据

CALIPSO卫星由美国和法国联合研发，其穿越赤道的时间分别约为1:30和13:30，重复周期为16天。CALIPSO的核心任务是提供全球连续观测的云和气溶胶数据集，从中评估气溶胶对气候强迫、云-气候反馈的直接和间接影响的不确定性。该卫星极大增强了研究者对云、气溶胶及其对地球环境

和气候系统影响的理解[20]。卫星搭载的主要仪器是CALIOP激光雷达[21],这是一种低轨双波长激光雷达,工作波长为532 nm和1 064 nm。其中,532 nm通道提供偏振测量信息,该通道能提升气溶胶类型的判别能力。CALIOP可在两个高度范围内对气溶胶测量:-0.5~8.2 km范围内具有0.333 km的水平分辨率和0.03 km的垂直分辨率;8.2~20.1 km范围内的水平分辨率为1 km,垂直分辨率为0.06 km。

本研究使用的是CALIPSO气溶胶剖面L3数据产品,版本号为V4.20。该数据产品提供经过质量控制的全球气溶胶剖面月平均值,并将数据分配于规则网格中。数据覆盖除南极洲外的纬度范围为85°S至85°N(分辨率为2°),经度范围为180°W至180°E(分辨率为5°),垂直海拔范围为-0.5~12 km(分辨率为0.06 km)[22]。此外,后续分析中使用了AOD和ALH(Aerosol Layer Height)参数,AOD和ALH分别表征了气溶胶的整体消光能力和等效高度,其定义如下[23]:

$$\mathrm{AOD}=\sum_{i=1}^{N}\mathrm{Ext}_i\cdot\Delta h_i \quad , \quad (1)$$

$$\mathrm{ALH}=\frac{\sum_{i=1}^{N}\mathrm{Ext}_i\cdot\Delta h_i\cdot h_i}{\sum_{i=1}^{N}\mathrm{Ext}_i\cdot\Delta h_i} \quad , \quad (2)$$

式(1)和(2)中Ext_i表示垂直方向上第i层的消光系数,h_i表示第i层的海拔高度,$\triangle h_i$表示对应高度第i层的空间垂直分辨率。

1.2　研究区域

本文选择中国大陆及临近区域作为研究区域,经度范围从73°E至135°E,纬度范围从18°N至55°N。使用的数据时间范围为2010年1月至2020年12月,共计22 096条网格数据。中国大陆地区气候多样且复杂,涵盖了温带、亚热带和热带气候带,使得该地区的气溶胶分布特征具有更加显著的空间差异[24]。另外,中国的气候受季风影响显著,冬季受西伯利亚冷空气的影响,夏季则受到来自南方的暖湿气流的影响,导致不同季节气溶胶的类型和AOD发生较大变化[25]。研究区域内的地貌差异,如高山、平原、沙漠和海洋的分布,也直接影响气溶胶的来源、运输和沉降过程,从而进一步增加了气溶胶的时空变异性。总的来说,该区域内的数据对研究气溶胶垂直分布的空间和季节变化规律有着重要意义。

1.3　聚类方法介绍

由于本文所使用的CALIPSO L3气溶胶剖面产品不包含先验知识的标签,因此采用聚类分析方法对气溶胶剖面进行分类。聚类分析是一种无监督学习方法,根据数据对象的相似性将其分组,从而使同一组内的数据相似度较高,而不同组间的数据差异显著[26-27]。尽管聚类方法的核心是基于数据相似性或距离进行分类,但不同方法在分类策略和依据上存在差异,这可能导致最终结果有所不同。聚类方法的灵活性和对复杂场景的适应能力,使其在数据挖掘和模式识别等领域得到了广泛应用[28-29]。接下来将简要介绍拟采用的聚类方法及其特点。

1.3.1　*K*-means聚类

K-means算法最初是一种用于信号处理的向量量化方法,由James MacQueen等人首次提出[30]。该算法的目标是通过最小化数据点到簇中心的距离(即簇内平方误差和)来实现数据的聚类。*K*-means的具体步骤如下:首先随机选择*K*个数据点作为初始簇中心;然后根据每个数据点与簇中心的距离,将数据点分配到最近的簇;最后重新计算每个簇的质心。上述过程将不断迭代,直到达到预设的最大迭代次数或质心位置不再发生变化。数据点到质心的距离定义如下:

$$J=\sum_{i=1}^{K}\sum_{x\in C_i}\left\|x-\mu_i\right\|^2 \quad , \quad (3)$$

式(3)中K表示簇的数量,x表示数据点,C_i为第i个簇中数据点的集合,μ_i为第i个簇的质心。*K*-means聚类算法的优点包括算法实现简单、收敛速度快等。然而,该算法的聚类结果对初始簇中心的选择敏感,不同的初始化状态可能导致该方法收敛于局部最优解。

1.3.2　GMM聚类

高斯混合模型聚类是一种基于模型的聚类方法,这类方法通常假设数据分布符合某种统计模型。GMM假设数据由有限多个高斯分布的线性组合构成,每个高斯分布代表一个未聚类的潜在簇[31]。这些分布的参数(如均值向量、协方差矩阵和混合系数)通过期望最大化算法(Expectation-Maximization Algorithm,EM)进行估计。EM算法通过在"期望步骤"和"最大化步骤"之间交替迭代,逐步优化参数,直至参数变化小于预设阈值[32]。在GMM中,均值向量和协方差矩阵定义每个高斯分布的形状和位置,而混合系数则表征该簇在整体数据

分布中的权重。GMM的概率密度函数可以表示为：

$$p(x)=\sum_{i=1}^{K}\pi_k\cdot N\left(x|\mu k,\sum k\right) \quad , \quad (4)$$

式(4)表示各个簇的高斯分布加权和，式中K为高斯分布的数量，π_k为高斯分布的混合系数。GMM聚类的优势在于提供软聚类结果，通过概率分布分配样本归属权重。相较于硬聚类方法(如K-means)，它能够灵活拟合不同方向和椭圆形状的数据分布，对复杂数据有更好的适应性。然而，GMM假设数据由高斯分布混合生成，对非高斯分布效果有限，且需预先指定簇数量，对初始化敏感。

1.3.3 谱聚类

谱聚类(Spectral Clustering)方法是一种基于图论的聚类算法，其核心思想是将数据点视为图的节点，通过构建相似度矩阵量化点间关系，再借助拉普拉斯矩阵的谱性质实现聚类[33]。谱聚类的主要步骤包括：构建相似度矩阵、计算拉普拉斯矩阵、进行特征分解以及应用聚类算法。具体而言，相似度矩阵用于表示数据点之间的相似性，其元素由高斯核函数定义，用于量化数据点x和y之间的相似程度。拉普拉斯矩阵则通过对相似度矩阵进行归一化处理，用以捕捉数据的局部几何特性。之后，通过对拉普拉斯矩阵进行特征分解，将数据映射到低维空间。高斯核函数的定义如下：

$$w=\exp\left(-\frac{\left\|x-y\right\|^2}{2\sigma}\right) \quad , \quad (5)$$

式(5)中σ表示高斯核的带宽，其决定了在计算数据点间相似度时权重的衰减速度。较小的σ值会导致相似度矩阵更加稀疏，而较大的σ则可能导致权重分布过于均匀，从而影响聚类结果的精确性。谱聚类方法能够处理非凸分布的数据，并且对噪声和异常数据表现出较强的抗干扰性。然而，谱聚类算法在构建相似度矩阵时计算复杂度较高，尤其在处理大规模数据时可能成为性能瓶颈。此外，谱聚类的效果对相似度矩阵的参数高度依赖，参数选择不当可能显著影响聚类结果的准确性。

1.4 数据预处理

根据气象条件，CALIOP L3产品分为四种子类型，即全天空数据(All_Sky)、无云数据(Cloud_Free)、多云天空不透明数据(Cloudy_Sky_Opaque)以及多云天空透明数据(Cloudy_Sky_Transparent)[34]。其中，全天空数据包含所有观测条件的数据，无云数据排除了云层对气溶胶的干扰，而多云天空数据进一步区分为不透明和透明两种类型，以表征云层的不同光学特性。无云产品在从气溶胶剖面中去除云层影响时，不可避免地会对气溶胶进行误判，从而减少气溶胶的判识量，可能进一步导致气溶胶消光系数的低估[35]。但这种误判带来影响远小于云对气溶胶的影响。另外，研究表明CALIPSO夜间数据的信噪比优于白天测量数据[36]，因此本文选择无云条件下的夜间产品作为本研究的研究数据。

为了更全面地表征垂直剖面信息并改进分类效果，除了原始的CALIPSO气溶胶消光系数剖面、AOD和ALH以外，还引入了其他特征变量。特征值的计算前需要对原始消光曲线进行归一化处理，将气溶胶消光系数曲线缩放到合适的范围，从而消除数值量级的影响。归一化过程如下：

$$x'=-\frac{x-\mathrm{mean}(x)}{\max(x)-\min(x)} \quad , \quad (6)$$

式(6)中x和x'分别代表归一化前后的消光系数。在此基础上，为了更全面地表征气溶胶消光系数随高度变化的特性，进一步引入了六个特征：均值、最大值、最小值、标准差、曲线下的面积(积分值)以及垂直变化率。这些特征能够从不同角度描述消光系数的分布特征和随高度的变化趋势。所有特征均经过式(6)标准化处理以消除量纲差异。在最终输入中增加6个垂直特征与2个时空标签，增加的数据维度有助于通过聚类方法同时识别气溶胶剖面的垂直结构与时空分布特征，为后续时空演变规律的分析奠定基础。

2 结果与分析

2.1 聚类数K选择

由于备选聚类方法的聚类效果受K值的选择影响较大，且不同聚类方法适用的量化评价指标各不相同，因此需要结合多种评估指标来确定最优的K值。肘部法则(Elbow Method)通过分析簇内平方和(Within-Cluster Sum of Squares, WCSS)随簇数K的变化趋势，确定聚类算法的最优簇数。WCSS定义为所有数据点到其所属簇中心的欧氏距离平方之和，其值随K增大单调递减，但下降速率在真实簇数附近会出现显著减缓，形成拐点(即“肘部”)[37]。本研究对标准化后的数据计算不同K值对应的WCSS绘制如图1(a)所示的关系曲线。肘部法则特别适用于以最小化簇内平方误差为目标的基于质心的聚类方法，例如K-Means和K-Medoids。通过观察曲线斜率变化可知，当K值从2增加到4时，WCSS

呈现快速下降趋势,表明聚类效果显著提升;而当K增至5时,WCSS的下降速度显著减缓;当K大于5时,WCSS的变化趋于平缓,说明增加聚类数对整体聚类效果的提升已有限。因此,可以确定K=5即为肘部位置。

对于GMM、谱聚类等,不基于质心的聚类方法,则采用轮廓系数(Silhouette Coefficient)来确定最优的K值[38]。轮廓系数是一种综合考虑簇内紧密度和簇间分离度的评价指标,其值分布在[-1,1]之间,其中值越大表示聚类效果越好。图1(b)展示了K值与轮廓系数的关系曲线。从图中可以看出,轮廓系数K值的变化呈现先升后降趋势。具体来说,当K值从2增加至4时,轮廓系数逐渐增大;而当K大于4时,轮廓系数开始下降。然而,由于轮廓系数在处理复杂数据集结构或簇形状不规则时可能不够稳健,其下降并不总是完全反映聚类质量的变化。此外,尽管K值为4和K值为5时轮廓系数的差异较小,但目视检查聚类结果显示,K值为5时的聚类效果更能准确捕捉气溶胶剖面的关键特征。因此,综合考虑肘部法则和轮廓系数曲线以及聚类目视效果,本研究最终将气溶胶剖面的聚类数K确定为5。

2.2　聚类方法选择

在确定了最佳聚类数K后,需进一步筛选适配气溶胶垂直分布特性的聚类算法。图2展示了三种方法的初步聚类结果。通过观察图2(a)和图2(b),可以发现GMM方法和K-means方法在捕捉气溶胶垂直分布的消光曲线形状和变化趋势方面表现良好。这两种方法生成的聚类结果能够清晰反映不同簇之间的垂直结构差异,为后续的分析提供了可靠基础。相较之下,从图2(c)可以看出,谱聚类方法在表征气溶胶消光曲线的垂直分布结构时能力较弱。具体来说,该方法生成的聚类曲线整体变化趋势相似,各簇之间的主要区别仅体现在AOD的大小上。此外,谱聚类生成的Cluster 5的消光曲线随高度变化幅度极小,几乎平行于x=0这条直线,表明该方法未能准确刻画消光曲线的垂直变化特征。这种局限性可能因为谱聚类方法对非线性和复杂数据结构的适应性不足,造成了图结构稀疏化,最终使拉普拉斯矩阵特征分解的信噪比降低。综合考虑聚类结果的适配性和研究目标,可得谱聚类方法无法满足本研究对气溶胶垂直分布结构的精确表征需求,因此将其排除。而对于GMM方法和K-means方法的取舍,则需要结合进一步的分析对两种方法的聚类效果进行比较。

为进一步确定在GMM方法和K-means方法之间的最佳选择,首先采用主成分分析(Principal Component Analysis,PCA)对高维气溶胶数据进行降维处理。PCA是一种常用的数据降维方法,通过线性变换将高维数据投影到低维空间中,同时保留数据的主要信息(如总方差)和特征分布[39]。此处PCA被用于将气溶胶数据降至三维,以便可视化比较聚类结果分布。图3(a)、(b)分别展示了GMM方法和K-means方法的三维聚类结果。需要注意的是,未降维数据中的某些簇在降维后的主成分方向上可能分布较为稀疏,即在低维投影中呈现位置较广或样本点集中于某一部分,导致视觉上数据点数量减少。从图中可以观察到,GMM方法的聚类结果在三维空间中分布较为均匀,各个簇之间的分离度较高。这表明,GMM能够更好地捕捉气溶胶消光曲线类型之间的差异性,清晰地区分不同的气溶胶垂直分布特征。而对于K-means方法,其聚类可视化结果显示出两个簇的样本数量占比过低(Cluster 4占比仅为0.96%,Cluster 5占比仅为0.77%),反映

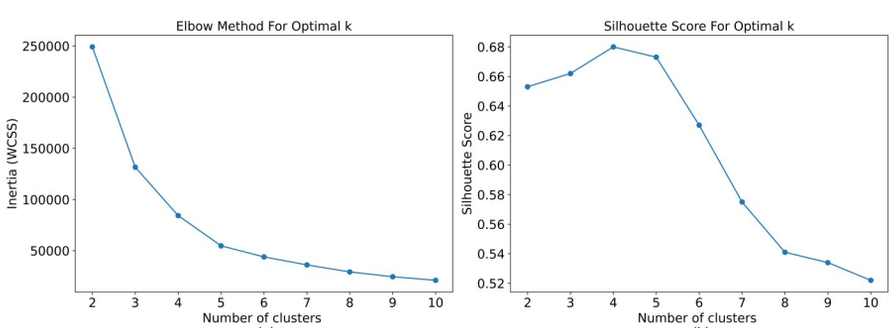

图1　利用肘部法则和轮廓系数确定最佳K值:(a)WCSS与聚类数K之间的关系;(b)轮廓系数与聚类数K之间的关系

Fig. 1　Determining the optimal K value using the Elbow Method and Silhouette Coefficient: (a) the relationship between WCSS and the number of clusters K; (b) the relationship between the Silhouette Coefficient and the number of clusters K

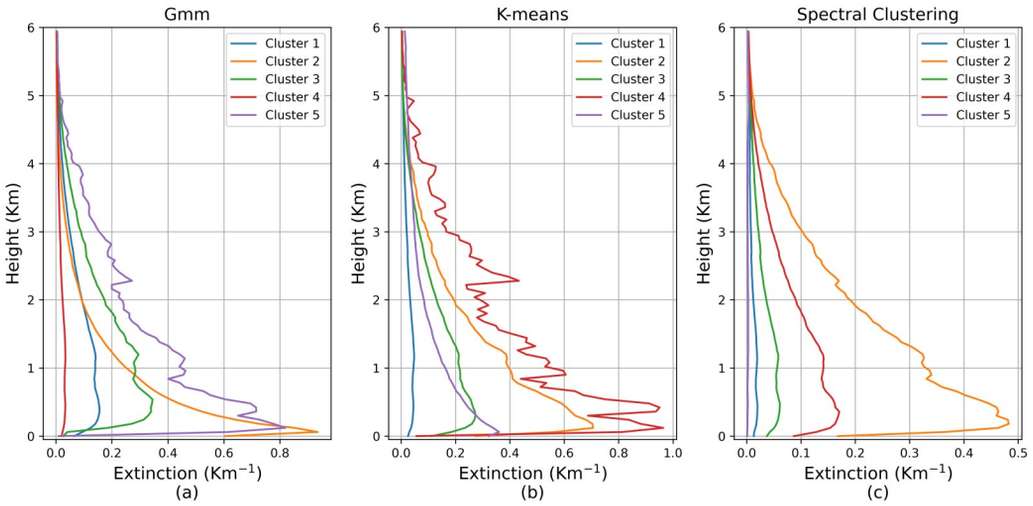

图2 三种待选聚类方法(GMM,*K*-means,Spectral Clustering)的聚类结果

Fig. 2 The clustering results of the three candidate clustering methods (GMM, *K*-means, and Spectral Clustering)

其未能形成有效簇结构。这是由于*K*-means方法对质心位置的过度依赖,难以适应数据复杂的非线性分布,从而导致部分簇无法有效代表气溶胶的实际特性。综合上述分析,GMM方法在本文研究的问题中表现出更强的分离能力和对气溶胶垂直分布的表征能力。因此,本研究最终选择GMM方法作为气溶胶垂直分布的聚类方法。

2.3 聚类结果

基于图2(a),根据消光曲线的垂直分布特征和污染情况对GMM的聚类结果进行了重新命名。聚类1和聚类3的消光曲线形态相似,分别由上半部分的指数衰减曲线和下半部分的接近均匀分布组成,因此将其分别命名为低污染组合型(Low-Pollution Composite Type,LPCT)和高污染组合型(High-Pollution Composite Type,HPCT)。聚类2的消光曲线除近地表外均呈现为典型的指数衰减形态,即气溶胶浓度随海拔升高快速降低,因此命名为指数衰减型(Exponential Decay Type,EDT)。聚类4的消光能力始终较低,且几乎不随高度变化,因此将其命名为低污染均匀型(Low-Pollution Uniform Type,LPUT)。聚类5消光曲线表现出显著的污染特征,同时随高度呈现剧烈的波动变化,因此将其命名为高污染振荡型(High-Pollution Oscillatory Type,HPOT)。

图4和图5分别展示了5种气溶胶剖面的详细垂直分布以及对应的AOD和ALH分布特征。由于存在极端异常值,图4中各子图的x轴范围根据数据分布特性分别进行了调整,整体未采用统一尺

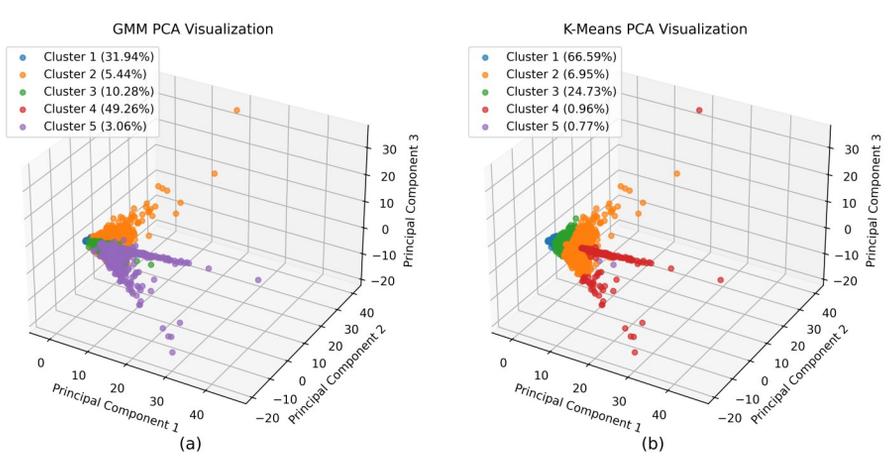

图3 经过PCA处理后的聚类方法的三维可视化数据图:(a)Gmm;(b)*K*-means

Fig. 3 Three-dimensional visualization of clustering methods after PCA processing: (a) for GMM; (b) for *K*-means

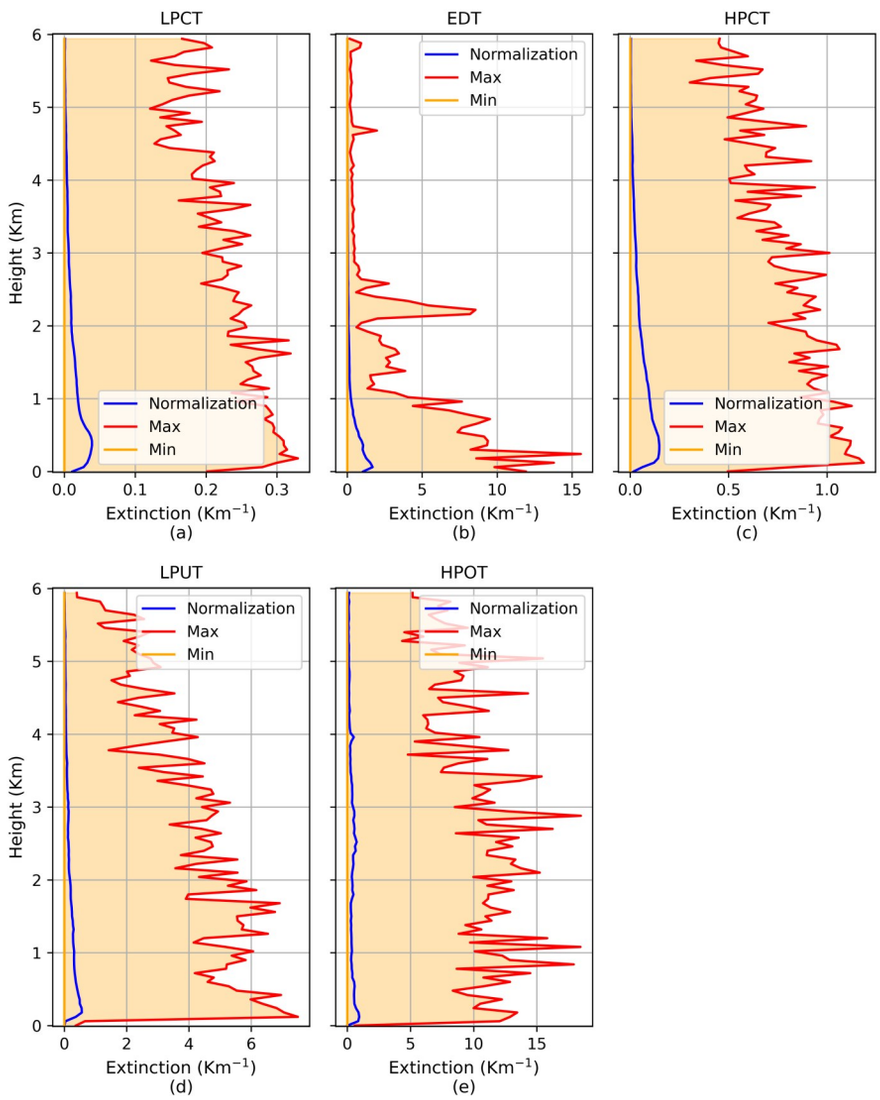

图4　由GMM聚焦分析后得到的5个气溶胶剖面的平均消光曲线以及最大、最小值曲线

Fig. 4　The average extinction curves, along with the maximum and minimum value curves, of the five aerosol profiles obtained through GMM-focused analysis

度。LPCT型和HPCT型剖面的消光最大值出现在0.2 km至0.5 km的高度范围内。当海拔高度低于0.2 km时,消光能力随高度降低快速减少;在0.2 km至1.1 km的高度范围内,消光变化趋于平缓;当高度超过1.1 km时,消光能力随高度增加逐渐减弱。两种剖面的AOD和ALH中位数分别为0.368和1.998、0.723和1.996。EDT型剖面的消光最大值出现在低于0.1 km的高度范围内。当海拔高度高于0.1 km时,消光呈指数衰减趋势,其AOD和ALH中位数分别为0.735和1.325。LPUT型剖面的消光能力始终较低(小于0.3),最大值出现在约1 km的高度处,其AOD和ALH的中位数分别为0.088和2.515。HPOT型剖面的消光随海拔高度呈现剧烈震荡,剖面通常出现多个峰值,其AOD和ALH中位数分别为1.227和1.653。此外,根据AOD的分布特征,可将五种剖面分为高污染型(HPCT、EDT、HPOT)和低污染型(LPCT、LPUT)。其中,高污染型剖面表现出较高的AOD值,反映了更高的气溶胶浓度。根据ALH的分布特征,可将五种剖面分为高海拔型(LPCT、HPCT、LPUT)和低海拔型(EDT、HPOT)。高海拔型剖面对应更高的ALH值,通常与更广的气溶胶传播范围相关,而低海拔型则集中于近地表区域。

与Wang等人使用模糊K-means方法得到的三种气溶胶剖面聚类结果相比[18],本文得到的聚类结果为五种典型消光剖面,聚类结果更加细化。每种

典型消光剖面的垂直分布特征也更加明显，能够更精确地反映气溶胶在不同高度上的分布规律。这是因为，一方面，本文在数据预处理过程中增加了气溶胶消光系数的高度变化特性，且选取了更适合数据特征的聚类方法；另一方面，相较于Wang等人的研究区域，本文的研究区域在地理分布、气候特点以及大气环境方面更为复杂，这也使得气溶胶垂直分布的聚类结果表现出更为多样化的特征。

2.4 气溶胶剖面季节分布特征

为了探究研究区域内气溶胶剖面的季节变化特征，将3月、4月、5月定义为春季；6月、7月、8月定义为夏季；9月、10月、11月定义为秋季；12月、1月、2月定义为冬季，并绘制了聚类随季节变化的结果，如图6所示。低污染类型的剖面（LPCT型和LPUT型）表现出显著的季节性变化。LPCT型剖面在温度较低的春季和冬季占比较大，分别为40.04%和34.69%；而LPUT型剖面在温度较高的夏季和秋季占比增多，分别为58.44%和54.48%。对于高污染类型的剖面，HPCT型的变化幅度最大，其占比从春季的12.17%下降至秋季的8.45%，下降了约3.72%。这种变化与季节性污染源（如春季的沙尘暴、冬季的燃煤取暖）及不同季节大气清除能力的差异有关[40]。相比之下，HPOT型剖面的变化幅度最小，其占比在各季节之间的差异仅为1.13%（最大为3.43%，最小为2.30%）。

图7（a~d）和（e~f）分别展示了五种剖面在不同季节的AOD和ALH分布特征。从图7（a~d）可以看出，低污染类型的LPCT和LPUT剖面的AOD分布几乎不受季节变化影响，始终保持在较小的范围内。这表明低污染类型的气溶胶浓度较低，主要受背景大气条件的控制，较少受到季节性污染源的影响。而高污染类型的HPCT、EDT和HPOT剖面的AOD分布在春季和夏季表现出较高值，在秋季和冬季则显著降低。具体而言，HPOT剖面的季节变化幅度最大，其AOD中位数从春季的1.38下降至秋季的1.09。对于ALH分布（图7（e~f）），五种剖面均表现出相似的季节性特征。在温度较低的秋季和冬季，各剖面的ALH值更接近地表，中位数集中在1.0 km以下；而在温度较高的春季和夏季，各剖面的ALH值随海拔升高呈现更广泛的分布范围，中位数可达1.5 km以上。这种气溶胶垂直分布的季节性变化反映了气溶胶传输机制的差异：冬季较稳定的大气条件限制了气溶胶的垂直扩散，而春夏季较强的对流活动促进了气溶胶粒子向高空的传播。

2.5 气溶胶剖面地域分布特征

如图8所示，此处选取了青藏高原（Qinghai-Tibet Plateau，QTP）、京津冀（Beijing-Tianjin-Hebei，BTH）和长三角（Yangtze River Delta，YRD）三个典型地区，研究气溶胶剖面的地域分布特征。从图中可以看出，三个地区的气溶胶剖面中低污染类型数量均占主导地位。在青藏高原地区，LPUT型剖面占比高达88.9%，这一特征与该地区较低的人类活动强度及稳定的大气条件有关[41]。京津冀地区的高污染型气溶胶剖面占比最大，其中LPCT型占比为43.88%，这可能与中国北方春季常见的沙尘天气以及冬季燃煤取暖排放有关[42]。京津冀地区和长三

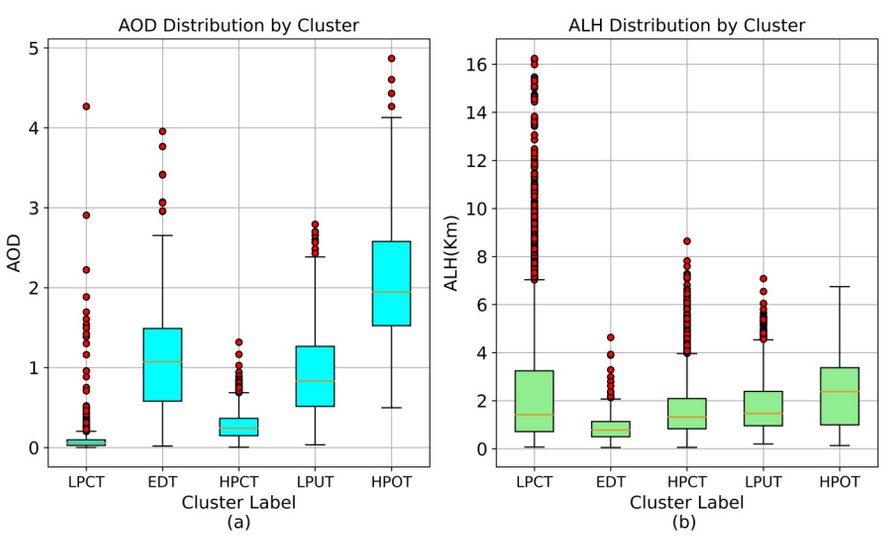

图5 由GMM获取的5个聚类结果：（a）AOD分布；（b）ALH分布

Fig. 5 The 5 clustering results obtained by GMM: (a) AOD distribution; (b) ALH distribution

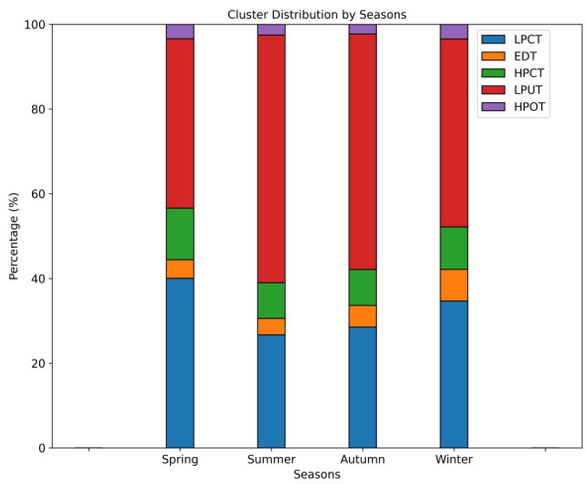

图6　气溶胶剖面聚类的季节分布特征

Fig. 6　Seasonal distribution characteristics of aerosol profile clustering

角地区的主要剖面类型分别为LPCT型(占比43.88%)和LPUT型(占比41.16%),而其他高污染型剖面的占比较小。EDT型剖面在这两个地区占比较高,分别位居第三,占比为18.62%和16.54%,这表明EDT型剖面在低海拔沿海地区的地域分布特征较为明显。总的来说,聚类分析结果揭示了五种气溶胶剖面在不同地区的明显地域分布特征。青藏高原地区以低污染剖面为主,显示了较少的人为干扰;而京津冀和长三角地区尽管也以低污染剖面类型为主导,但仍存在约30%的高污染型剖面分布。

图9展示了五种剖面在三个典型地区的AOD和ALH分布。从图9(a~c)可以看出,所有剖面类型在青藏高原地区均呈现较低的污染水平。其中,HPOT型剖面在京津冀地区的AOD波动较大,去除异常值后该地区此类剖面AOD最大值与最小值的差值为1.78。从图9(d~f)可以观察到,ALH在青藏高原地区展现出明显的地域分布特征。与京津冀和长三角地区相比,青藏高原的典型剖面(除EDT型外)的ALH中位数平均增加了2.81 km。

造成这一差异的原因与不同地区的地形、气象条件和污染排放源的差异密切相关。青藏高原地区地势较高,地形复杂,常年受高原季风影响,使得污染物不易在低层大气中积累[43]。这种气象特性导致气溶胶层高度更高,同时AOD值较低。此外,青藏高原地区的人为污染源较少,以自然来源的气溶胶为主。因此,青藏高原地区的气溶胶剖面类型的污染水平整体较低。相比之下,京津冀和长三角地区是典型的工业化和城市化区域,气溶胶来源更为复杂,包括工业排放、交通污染和生物质燃烧等。这些人为污染源的高排放量导致AOD值显著升高。

2.6　气溶胶类型分布特征

为了探究气溶胶类型的分布特征,此处分析了CALIOP L3产品提供的七种气溶胶亚型(包括清洁海洋、高烟、清洁大陆、沙尘海洋等)与季节之间的关系,结果如图10所示。从图中可见,占比最多的气溶胶亚型为污染尘埃型(Polluted Dust)、尘埃型(Dust)以及污染大陆/烟雾型(Polluted Continental/Smoke)。进一步分析显示,经过聚类得到的五种典型剖面的气溶胶亚型组分在秋冬季节温度较低时

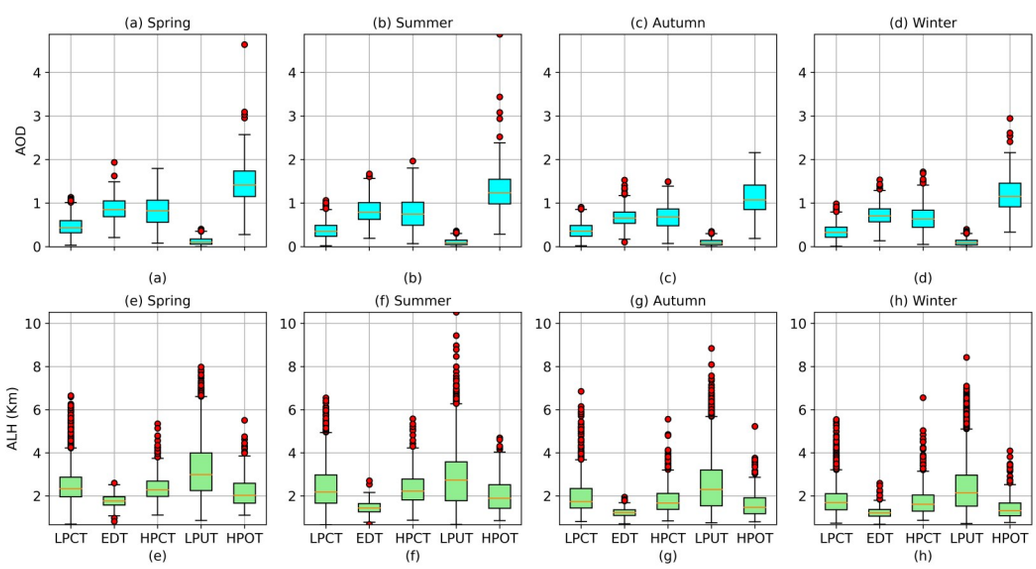

图7　五种剖面在不同季节的分布:(a~d)AOD分布;(e~h)ALH分布

Fig. 7　Seasonal distribution for the five aerosol profiles; (a~d) AOD distribution; (e~h) ALH distribution

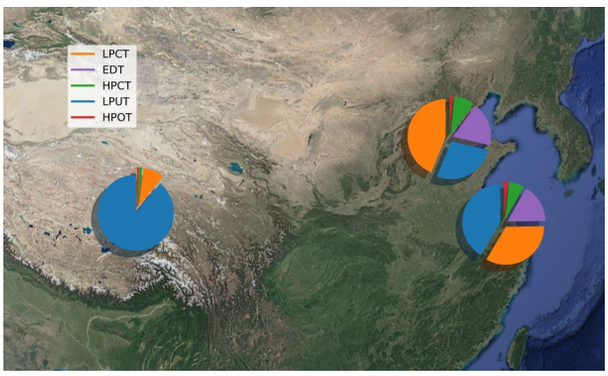

图8 气溶胶剖面聚类在中国地区的地域分布特征

Fig. 8 Geographical distribution characteristics of aerosol profile clustering in China

较为固定,其中污染尘埃型气溶胶的占比最多。具体而言,变化幅度最大的是EDT型剖面中的污染尘埃型气溶胶,其占比从秋季的29.01%下降至冬季的22.68%,表明其组分具有较高的季节稳定性。在春夏季节温度较高时,各类剖面的气溶胶组成变化较为复杂,尤其是EDT型剖面的污染大陆/烟雾型气溶胶,其占比从春季的0.74%增加到夏季的28.22%。造成这种现象的原因与春夏季节气溶胶源的多样性和大气条件的复杂性有关。另外,LPUT型剖面的组分占比较为稳定,这是由于该剖面主要出现在污染程度较低的地区,其气溶胶来源和大气环境特征受到季节变化的影响较小。

3 结论

本文利用三种聚类算法对2010年至2020年的CALIOP L3气溶胶剖面数据集进行了空间聚类分析,根据气溶胶剖面的垂直分布特征,将其划分为5类具有代表性的气溶胶剖面类型。本文的主要结论如下:

(1)GMM聚类方法能够较好地表征气溶胶剖面的垂直分布特征。通过GMM聚类得到的5个簇清晰地反映了气溶胶垂直分布的典型特征,聚类结果具体包括:低污染组合型(LPCT)、高污染组合型(HPCT)、指数衰减型(EDT)、低污染均匀型(LPUT)和高污染振荡型(HPOT)。

(2)气溶胶剖面的季节性变化表现出显著的特征。低污染类型的剖面(LPCT型和LPUT型)在春季和冬季占比较大,分别为40.04%和34.69%,而在夏季和秋季则有所增加,特别是LPUT型剖面,占比在夏季和秋季分别达到58.44%和54.48%。高污染类型的剖面中,HPCT型的季节性变化较为明

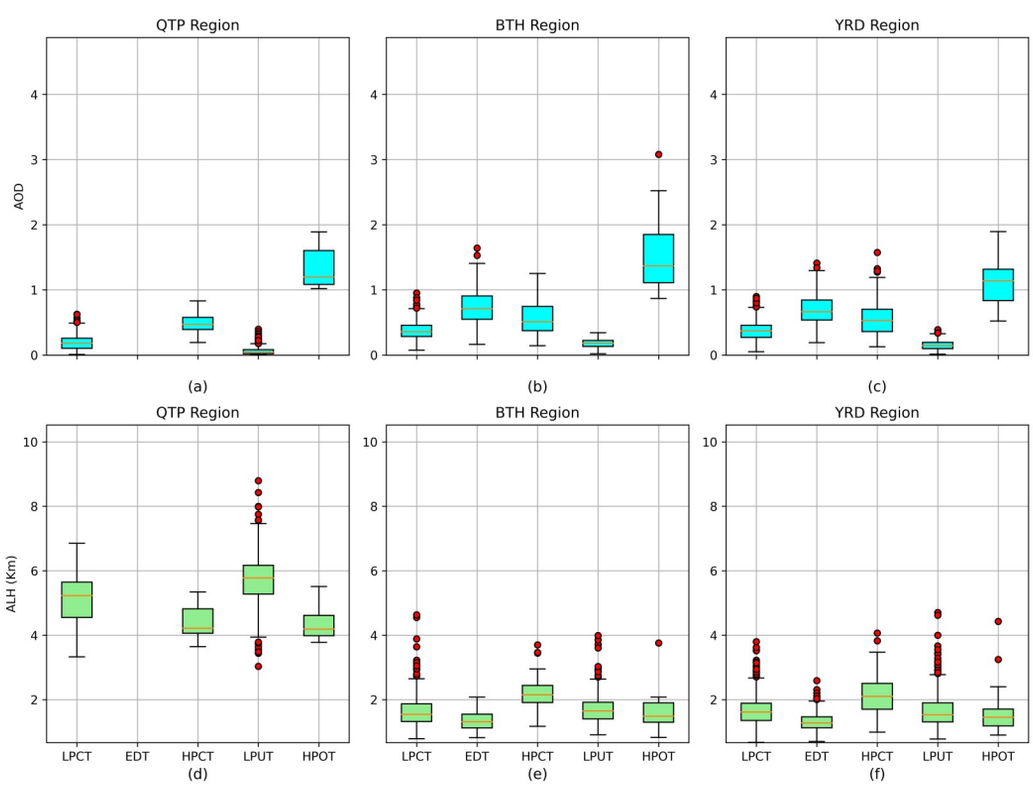

图9 五种剖面在典型地域上的分布:(a~c)AOD分布;(d~f)ALH分布

Fig. 9 The distribution for the five profiles in typical regions: (a~c) AOD distribution; (d~f) ALH distribution

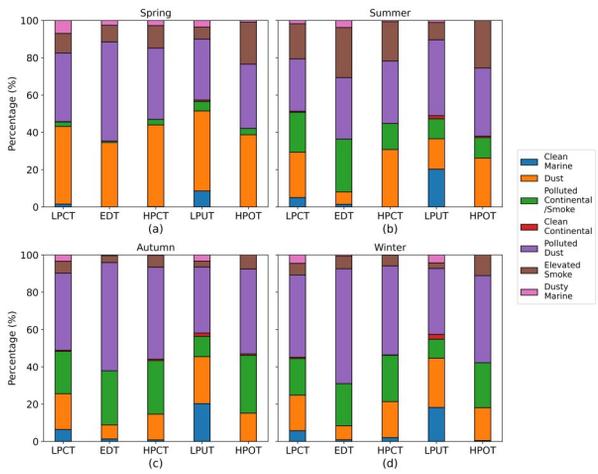

图10　五种消光剖面在不同季节的气溶胶类型组成

Fig. 10　The aerosol type composition of the five extinction profiles in different seasons

显,其占比从春季的12.17%下降至秋季的8.45%,下降幅度约为3.72%。

(3)气溶胶剖面的类型在不同地域呈现出显著的分布差异。青藏高原以低污染类型为主,LPUT型剖面占比高达88.9%;京津冀地区则以高污染类型为主,LPCT型占比为43.88%。长三角地区主要为LPCT型和LPUT型剖面,EDT型剖面在京津冀和长三角地区占比较高。

(4)五种消光剖面在不同季节的气溶胶类型组成稳定性不同。温度较低秋冬季节,五种典型剖面的气溶胶组成较为固定;而在温度较高的春夏季,不同剖面的气溶胶组分变化复杂。

本文的主要贡献包括两个方面:一是为辐射传输模拟模型提供了更精细的气溶胶垂直分布模型,该模型有助于优化辐射传输计算并提升其精度;二是验证了在基于机器学习进行气溶胶垂直分布反演中引入时间和区域标签等先验信息的可行性与有效性。

未来的研究可以从时间和空间两个维度对数据量进行扩充,通过引入更长时间跨度和更大范围的数据,捕捉更全面的气溶胶垂直分布动态特征,提升研究结果的普适性,以支持更大范围的气溶胶分布特性研究和应用。另外,也可以结合L2气溶胶剖面数据获取垂直特征更为突出的气溶胶剖面结构。

References

[1] Calvo A I, Alves C, Castro A, et al. Research on aerosol sources and chemical composition: Past, current and emerging issues[J]. Atmospheric Research, 2013, 120: 1-28.

[2] Colbeck I, Lazaridis M. Aerosols and environmental pollution[J]. Naturwissenschaften, 2010, 97(2): 117-131.

[3] Song C, He J, Wu L, et al. Health burden attributable to ambient PM2. 5 in China[J]. Environmental pollution, 2017, 223: 575-586.

[4] Hoek G, Krishnan R M, Beelen R, et al. Long-term air pollution exposure and cardio- respiratory mortality: a review[J]. Environmental Health, 2013, 12(1): 43.

[5] Mehta M, Singh N, Anshumali. Global trends of columnar and vertically distributed properties of aerosols with emphasis on dust, polluted dust and smoke - inferences from 10-year long CALIOP observations[J]. Remote Sensing of Environment, 2018, 208: 120-132.

[6] Babu S S, Moorthy K K, Manchanda R K, et al. Free tropospheric black carbon aerosol measurements using high altitude balloon: Do BC layers build "their own homes" up in the atmosphere?[J]. Geophysical Research Letters, 2011, 38(8): L08803-1 - L08803-6.

[7] Ge C, Wang J, Reid J S. Mesoscale modeling of smoke transport over the Southeast Asian Maritime Continent: coupling of smoke direct radiative effect below and above the low-level clouds[J]. Atmospheric Chemistry and Physics, 2014, 14(1): 159-174.

[8] Remote sensing of surface visibility from space: A look at the United States East Coast[J]. Atmospheric Environment, 2013, 81: 136-147.

[9] Wang Y, Sun X, Huang H, et al. Study on influencing factors of the information content of satellite remote-sensing aerosol vertical profiles using oxygen A-band[J]. Remote Sensing, 2023, 15(4): 948.

[10] Koffi B, Schulz M, Bréon F M, et al. Evaluation of the aerosol vertical distribution in global aerosol models through comparison against CALIOP measurements: AeroCom phase II results[J]. Journal of Geophysical Research: Atmospheres, 2016, 121(12): 7254-7283.

[11] Kessner A L, Wang J, Levy R C, et al. Remote sensing of surface visibility from space: A look at the United States East Coast[J]. Atmospheric Environment, 2013, 81: 136-147.

[12] Wang T, Han Y, Huang J, et al. Climatology of Dust-Forced Radiative Heating Over the Tibetan Plateau and Its Surroundings[J]. Journal of Geophysical Research: Atmospheres, 2020, 125(17): e2020JD032942.

[13] Liu Y, Zhu Q, Huang J, et al. Impact of dust-polluted convective clouds over the Tibetan Plateau on downstream precipitation[J]. Atmospheric Environment, 2019, 209: 67-77.

[14] Winker D M, Vaughan M A, Omar A, et al. Overview of the CALIPSO mission and CALIOP data processing algorithms[J]. Journal of Atmospheric and Oceanic Technology, 2009, 26(11): 2310-2323.

[15] Yu H, Chin M, Winker D M, et al. Global view of aerosol vertical distributions from CALIPSO lidar measurements and GOCART simulations: Regional and seasonal variations[J]. Journal of Geophysical Research: Atmospheres, 2010, 115(D4): D00H30-1-D00H30-19.

[16] Winker D M, Tackett J L, Getzewich B J, et al. The glob-

al 3-D distribution of tropospheric aerosols as characterized by CALIOP[J]. Atmospheric Chemistry and Physics, 2013, 13(6): 3345-3361.
[17] Fan Y, Sun X, Huang H, et al. The primary aerosol models and distribution characteristics over China based on the AERONET data[J]. Journal of Quantitative Spectroscopy and Radiative Transfer, 2021, 275: 107888.
[18] Wang L, Lyu B, Bai Y. Global aerosol vertical structure analysis by clustering gridded CALIOP aerosol profiles with fuzzy k-means[J]. Science of the Total Environment, 2021, 761: 144076.
[19] Zeng S, Vaughan M, Liu Z, et al. Application of high-dimensional fuzzy k-means cluster analysis to CALIOP/CALIPSO version 4.1 cloud-aerosol discrimination[J]. Atmospheric Measurement Techniques, 2019, 12(4): 2261-2285.
[20] Crisp D. Measuring atmospheric carbon dioxide from space with the Orbiting Carbon Observatory-2 (OCO-2)[C]. Earth observing systems xx: Vol. 9607. SPIE, 2015: 960702.
[21] Hunt W H, Winker D M, Vaughan M A, et al. CALIPSO lidar description and performance assessment[J]. Journal of Atmospheric and Oceanic Technology, 2009, 26(7): 1214-1228.
[22] Tackett J L, Winker D M, Getzewich B J, et al. CALIPSO lidar level 3 aerosol profile product: Version 3 algorithm design[J]. Atmospheric Measurement Techniques, 2018, 11(7): 4129-4152.
[23] Lu Z, Wang J, Chen X, et al. First Mapping of Monthly and Diurnal Climatology of Saharan Dust Layer Height Over the Atlantic Ocean From EPIC/DSCOVR in Deep Space[J]. Geophysical Research Letters, 2023, 50(5): e2022GL102552.
[24] Lei Y, Zhang Q, He K B, et al. Primary anthropogenic aerosol emission trends for China, 1990 - 2005[J]. Atmospheric Chemistry and Physics, 2011, 11(3): 931-954.
[25] Wu G, Li Z, Fu C, et al. Advances in studying interactions between aerosols and monsoon in China[J]. Science China Earth Sciences, 2016, 59(1): 1-16.
[26] Aggarwal C C. An introduction to cluster analysis[M]. Data clustering. Chapman and Hall/CRC, 2018: 1-28.
[27] Frades I, Matthiesen R. Overview on Techniques in Cluster Analysis[M]. Matthiesen R. Bioinformatics Methods in Clinical Research: Vol. 593. Totowa, NJ: Humana Press, 2010: 81-107.
[28] Miu Y W. The analysis of data based on the hierarchical clustering[D]. Anhui University, 2013.
缪元武. 基于层次聚类的数据分析[D]. 安徽大学, 2013.
[29] Saha R, Tariq M T, Hadi M, et al. Pattern Recognition Using Clustering Analysis to Support Transportation System Management, Operations, and Modeling[J]. Journal of Advanced Transportation, 2019, 2019 (Pt. 5): 1628417.1- 1628417.12.
[30] MacQueen J. Some methods for classification and analysis of multivariate observations[C]. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: Vol. 1. Oakland, CA, USA, 1967: 281-297.
[31] Rasmussen C. The Infinite Gaussian Mixture Model[C]. Advances in Neural Information Processing Systems: Vol. 12. MIT Press, 1999.
[32] Pfeifer T, Protzel P. Expectation-Maximization for Adaptive Mixture Models in Graph Optimization[C]. 2019 International Conference on Robotics and Automation (ICRA). 2019: 3151-3157.
[33] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.
[34] Nan Y, Wang Y. De-coupling interannual variations of vertical dust extinction over the Taklimakan Desert during 2007-2016 using CALIOP[J]. Science of the Total Environment, 2018, 633: 608-617.
[35] Pan H, Huang J, Kumar K R, et al. The CALIPSO retrieved spatiotemporal and vertical distributions of AOD and extinction coefficient for different aerosol types during 2007 - 2019: A recent perspective over global and regional scales [J]. Atmospheric Environment, 2022, 274: 118986.
[36] Kar J, Vaughan M A, Lee K P, et al. CALIPSO lidar calibration at 532 nm: version 4 nighttime algorithm[J]. Atmospheric Measurement Techniques, 2018, 11(3): 1459-1479.
[37] Ketchen Jr. D J, Shook C L. The application of cluster analysis in strategic management research: an analysisand critique [J]. Strategic Management Journal, 1996, 17(6): 441-458.
[38] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Journal of Intelligent Information Systems, 2001, 17(2/3): 107-145.
[39] Abdi H, Williams L J. Principal component analysis[J]. WIREs Computational Statistics, 2010, 2(4): 433-459.
[40] Huang K, Zhuang G, Li J, et al. Mixing of Asian dust with pollution aerosol and the transformation of aerosol components during the dust storm over China in spring 2007[J]. Journal of Geophysical Research: Atmospheres, 2010, 115(D7): 2009JD013145.
[41] Zhang X, Xu J, Kang S, et al. Chemical characterization and sources of submicron aerosols in the northeastern Qinghai - Tibet Plateau: insights from high-resolution mass spectrometry [J]. Atmospheric Chemistry and Physics, 2019, 19(11): 7897-7911.
[42] Kong F. Spatial and temporal evolution characteristics of days of disastrous dust weather in China from 1961 to 2017[J]. J. Arid. Land Resour. Environ., 2020, 34: 116-123.
[43] Huang L, Chen J, Yang K, et al. The northern boundary of the Asian summer monsoon and division of westerlies and monsoon regimes over the Tibetan Plateau in present-day[J]. Science China Earth Sciences, 2023, 66(4): 882-893.