一种光谱特征增强驱动的机器学习地基红外高光谱云检测方法*

王越[1)2)3)] 叶函函[2)3)†] 熊伟[1)2)3)‡] 王先华[2)3)] 施海亮[2)3)]
李超[4)] 程晨[2)3)] 吴时超[2)3)]

1) (中国科学技术大学环境科学与光电技术学院, 合肥 230026)

2) (中国科学院合肥物质科学研究院, 安徽光学精密机械研究所, 合肥 230031)

3) (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, 合肥 230031)

4) (中国科学技术大学地球和空间科学学院, 合肥 230026)

(2025 年 7 月 23 日收到; 2025 年 8 月 11 日收到修改稿)

云是地基红外高光谱仪器探测大气的重要干扰源, 有效云检测不可或缺. 水汽干扰和高云识别精度低是云检测面临的两个关键挑战. 本文利用大气红外光谱探测仪 (ASSIST) 在云南丽江、西藏自治区墨脱和西藏自治区日土的观测数据, 分析了晴空和有云条件下的光谱特征差异, 并据此提出了一种光谱特征增强的机器学习云检测方法. 结合同步观测的激光雷达、气象站及全天空成像仪数据, 系统评估了该方法在不同相对湿度 (RH) 和不同云底高度 (CBH) 条件下的检测性能. 实验结果表明: 该方法与激光雷达检测结果的一致性高达 97.61%. 在不同 RH 条件下, 该方法精度均优于使用原始光谱特征的方法, 尤其在 RH $>$ 70% 时, 对晴空光谱的识别精度提升明显, 从 86.01% 提高至 91.89%. 同样, 在不同 CBH 条件下, 新方法也展现出优于使用原始光谱特征方法的性能, 特别在识别 3 km $<$ CBH $\leqslant$ 5 km 的中云和 CBH $>$ 5 km 的高云时, 精度提升尤为明显. 当 3 km $<$ CBH $\leqslant$ 5 km 时, 精度从 95.45% 提升至 98.64%; 当 CBH $>$ 5 km 时, 精度从 87.5% 提升至 91.67%.

关键词: 地基红外高光谱, 遥感, 机器学习, 云检测

PACS: 02.70.Hm, 07.05.Mh, 07.57.Ty, 42.68.Ge **DOI:** 10.7498/aps.74.20250982

CSTR: 32037.14.aps.74.20250982

1 引 言

高光谱遥感能以高光谱分辨率获取数千个连续的光谱通道[1], 可以对地球大气的特征和细节进行更深入研究[2], 因此被广泛应用在气象领域中. 卫星上的高光谱红外探测仪可以定期提供覆盖全球的数据, 包括大气温湿度、温室气体、云和其他近地表特征的信息. 这些探测仪包括温室气体极轨干涉探测仪 (IMG)[3]、Aqua 上的光栅式大气红外探测仪 (AIRS)[4]、欧洲气象业务 (MetOp) 卫星上的红外大气探测仪 (IASI)[5], 以及 Suomi 国家极地轨道合作组织上的跨轨道红外探测仪 (CrIS)[6]. 此外, 中国也开发了自己的红外高光谱探测器, 如搭载在风云 3D(FY-3D) 卫星上的高光谱红外大气探测仪 (HIRAS)[7] 和风云 4A(FY-4A) 卫星上的地球静止干涉红外探测器 (GIIRS)[8]. 星载红外高光谱仪具有较宽的空间覆盖范围和较高空间分辨率,

* 国家重点研发计划 (批准号: 2022YFB3901804) 和安徽省自然科学基金 (批准号: 2408055UQ003) 资助的课题.

† 通信作者. E-mail: yehanhan@aiofm.ac.cn

‡ 通信作者. E-mail: frank@aiofm.ac.cn

但是在地表附近受影响显著. 相反, 地基红外高光谱仪能以更高时间分辨率接收大气下行辐射, 且受地表影响较小.

目前, 接收大气下行辐射的地基红外高光谱仪主要有部署在大气辐射测量 (ARM) 计划中的大气发射辐射干涉仪 (AERI)[9] 和 LR Tech 公司开发的大气红外光谱探测仪 (ASSIST)[10]. AERI 光谱范围覆盖 520—3000 cm^{-1} 的范围, 光谱分辨率为 1 cm^{-1}, 时间分辨率为 8 min, 可以在白天和夜晚连续测量低层大气中多种大气成分如温湿度 [11]、温室气体 [12]、气溶胶 [13]、云 [14] 等. 相比 AERI, ASSIST 具有更高的时间分辨率 (2 min), 其观测光谱覆盖了 520—3300 cm^{-1} 范围, 也被应用于多种大气成分的测量 [15,16]. 然而, 云在红外波段表现出强烈的吸收和散射效应, 对红外辐射有显著影响, 其影响通常远超过大气温度和成分分布不确定性引入的辐射扰动 [17]. 更重要的是, 数值天气预报 (NWP) 模型通常仅同化红外高光谱探测仪的晴空测量值, 因为很难用正向辐射模型精确模拟云污染条件下的辐射 [18]. 因此, 在应用 ASSIST, AERI 等地基红外高光谱仪器探测大气参数 (如温湿廓线) 时, 云检测是一个至关重要且不可或缺的步骤.

地基红外高光谱仪测量的大气下行辐射数据中包含了能够区分晴空和有云条件的信息, 但其光谱辐亮度值会受大气显著影响 (特别是温度、水汽). 水汽在红外波段的强吸收特性, 对云信号的探测构成了严重干扰 [19]. 传统的阈值云检测方法在应用时存在显著局限性, 主要在于其设定的阈值通常难以适应不同地点和动态变化的大气条件 [20]. Cho 等 [21] 使用 760—1000 cm^{-1} 辐射数据的 60% 作为阈值来判断 AERI 观测时的多云条件. 他们进一步尝试选取了四季典型无云数据来定义无云阈值, 但仍未能有效地解决阈值方法固有的普适性问题. 相比之下, 机器学习方法具备自动学习复杂非线性特征的能力, 有望实现更高精度、更强鲁棒性及更高效自动化的云检测, 因而被越来越多地应用于此领域, 常见模型包括随机森林 (RF)[22]、逻辑回归 (LR)[23]、轻量级梯度提升树模型 (LightGBM)[24] 和支持向量机 (SVM)[25] 等. 刘磊等 [25] 利用 AERI 在南大平原 (SGP)、北坡阿拉斯加 (NSA) 和南极辐射实验 (AWARE) 站点的观测数据, 基于 AERI 光谱数据中特定波段的特征信息, 提出了一种基于 SVM 模型的云检测方法, 并以云高仪的云检测结果作为基准进行了精度验证. 结果表明, 该算法在各站点的云检测结果与云高仪的一致性约为 93%. 然而, 该方法对于高云 (特别是薄卷云) 检测的一致性会明显降低.

为了进一步研究不同相对湿度 (RH) 以及不同云底高度 (CBH) 对地基红外高光谱辐射数据云检测的影响, 本文提出了一种光谱特征增强驱动的机器学习云检测方法. 基于该方法, 利用 ASSIST 在海拔、水汽及云高差异显著的云南丽江高美古天文台、西藏自治区墨脱气象观测站和西藏自治区日土县阿里荒漠环境综合观测站三地的观测数据, 分析了其光谱特征, 实现了云检测, 并与同步激光雷达观测结果进行了精度验证. 该方法有效地提升了地基红外高光谱辐射数据云检测的自动化程度与精度水平, 为后续的辐射传输模拟、遥感参数反演及数值天气预报 (NWP) 模型同化等应用提供了更高质量的基础数据支撑.

2 数 据

2.1 仪器

大气下行红外高光谱数据来源于 ASSIST. ASSIST 是专为满足美国国家核安全局 (NNSA) 的技术要求而开发的, 由 LR TECH 设计和制造 [10]. 作为美国能源部 (DOE) 推动的大气辐射测量 (ARM) 计划的一部分, 该仪器主要用于测量大气上行和下行辐射以及地面观测数据的验证. ASSIST 的核心由一个配备中波红外 (InSb) 和长波红外 (MCT) 探测器的干涉仪组成, 能够自动垂直观测 520—3300 cm^{-1} (3—19.2 μm) 的光谱范围下行红外辐射. 该仪器的光谱分辨率为 1 cm^{-1} (切趾后), 时间分辨率为 2 min, 视场角为 46 mrad, 最大光程差为 1.037 cm[26]. 本文使用 ASSIST 在云南丽江高美古 (26.695°N, 100.029°E)、西藏自治区墨脱气象观测站 (29.328°N, 95.293°E) 和西藏自治区日土县阿里荒漠环境综合观测站 (33.390°N, 79.703°E) 观测的数据, 为云检测算法提供晴空和多云样本.

同步观测的激光雷达数据被用作验证云检测算法精度的真值. 激光雷达设备采用的是型号为 LVIS-T100 的米散射微脉冲激光雷达 (MPL)[27], 它根据大气对激光的弹性散射、消光等物理效应, 通过探测大气气溶胶和云的激光后向散射回波,

实现对几公里乃至十几公里范围内的大气环境进行实时快速监测. 为了直观地判断 ASSIST 观测视场内云的存在情况, 同步配备了锦州阳光气象科技有限公司开发的型号为 TBK 11 的全天空成像仪进行实时拍摄. 该设备可以在无遮阳情况下 (即完全暴露于日光下) 自动清晰记录全天空云分布图像. 此外, 为了量化不同水汽条件对云检测的影响, 同步观测实验还部署了同样来自锦州阳光气象科技有限公司开发的 TRM-ZS2 型的高精度自动气象站, 用于实时监测环境温湿度、风速、风向等关键气象要素. 实验观测设备情况如图 1 所示. 为了确保观测数据时间一致性, 激光雷达、全天空成像仪及自动气象站的时间分辨率均设置为 1 min, 与 ASSIST 观测时间精确匹配.

图 1 实验配套设备

Fig. 1. Deployed instruments.

2.2 数据预处理

根据 ASSIST 在三个地点观测期间测量的多条异常光谱数据, 将测得的异常值分为 4 种情况, 如图 2 所示. 图 2(a) 和图 2(b) 出现的错误光谱是由于仪器开始测量时, 黑体不稳定导致的噪声过大造成的. 图 2(c) 出现的类似平滑曲线的光谱是由于仪器观测舱口盖子关闭, 使得无法接收大气的下行热辐射. 图 2(d) 展示的光谱出现多个负值, 这是测量不准确的表现. 图 2(e) 和图 2(f) 显示了两个典型的理想的无噪声的例子 (晴空和多云条件下). 根据这 4 种异常情况, 剔除 ASSIST 观测的错误光谱.

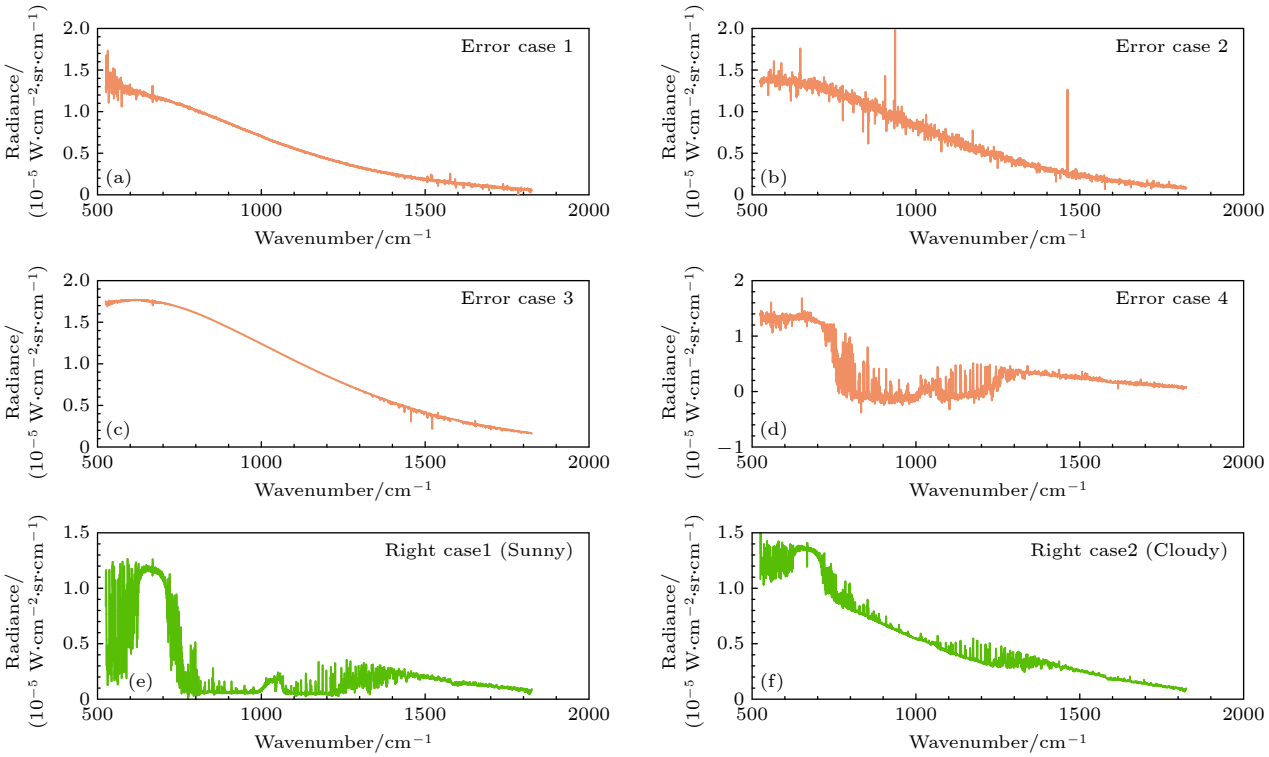

图 2 ASSIST 测得的几种错误光谱 (橙色线条) 和正确光谱 (绿色线条) (a), (b), (c), (d) 分别代表采集的 4 种错误光谱; (e), (f) 分别代表晴空和多云采集的正确光谱

Fig. 2. Several erroneous spectra (orange lines) and correct spectra (green lines) measured by ASSIST: (a), (b), (c), (d) respectively represent the four types of erroneous spectra collected; (e), (f) respectively represent the correct spectra collected under clear sky and cloudy conditions.

2.3 多云和晴空数据集

经过数据质量控制后, 建立多云和晴空光谱训练集和测试集. 首先, 根据全天空成像仪拍摄的晴空和多云的时刻查找 ASSIST 相同的时刻. 接着, 根据相同时刻激光雷达提供的云底高度数据进一步筛选出 ASSIST 多云和晴空光谱. 三个地点筛选出的晴空和多云样本情况如表 1 所列.

表 1 三个地点的晴空和多云样本数量

Table 1. The number of clear-sky and cloudy samples at three locations.

地点	晴空样本	多云样本	海拔/km	观测时间
丽江高美古天文台	3357	2826	3.23	2024.03.20—2024.05.04
墨脱气象观测站	1584	3641	0.76	2024.11.29—2024.12.19 2025.03.15—2025.03.28
日土阿里荒漠环境综合观测站	4052	1543	4.23	2025.05.27—2025.06.15
总计	8993	8010		

3 方 法

3.1 多云和晴空光谱特征分析

根据筛选出的晴空和多云样本, 在刘磊等[25]研究的基础上, 进一步总结了区分有云和晴空场景的特点. 图 3 展示了 ASSIST 在三个地点晴空和多云条件下的实测光谱. 可以明显地看出多云光谱和晴空光谱在 740—1200 cm^{-1} 波段存在显著的差异. 当有云存在时, 云会向外发射辐射, 导致该波段的辐亮度明显增大. 图 3 中浅紫色光谱区域展示了使用刘磊等的方法选择出的晴空和多云光谱特征, 具体特征如表 2 所列的前 12 个特征.

由于水汽分子对红外辐射有很强的衰减作用, 对云检测造成了较大干扰. 我们以 11 μm 波段中 4 个无气体和气溶胶吸收的通道 (包括 925.8524 cm^{-1}, 948.9987 cm^{-1}, 951.892 cm^{-1} 和 962.5007 cm^{-1} (绿色三角形)) 的光谱辐亮度以及它们与其相邻云和水汽发射能量较强的通道 (包括 925.3702 cm^{-1}, 948.5165 cm^{-1}, 951.4098 cm^{-1} 和 962.0185 cm^{-1} (红色圆圈)) 光谱辐亮度的比值, 作为新增的特征来区分有云和晴空场景, 具体特征见表 2 所列的后 8 个特征. 这是因为这 4 个无气体和气溶胶吸收的通道的气体光学厚度较低, 使得它们对上方云的发射非常敏感. 此外, 云在空间上很少是均匀的, 即使非常小的冷凝水数量也可以大大增加在 11 μm 窗口中观察到的辐射[28].

3.2 基于机器学习的云检测算法

利用 ASSIST 测量数据进行云检测可以看作是一个分类问题. 本文采用支持向量机 (SVM) 算法进行云检测. SVM 是解决小样本、非线性、高维问题的有效工具, 它的核心思想是利用分类超平面作为判别的基础, 达到最大程度的分类[29]. 该算法的主要流程如图 4 所示.

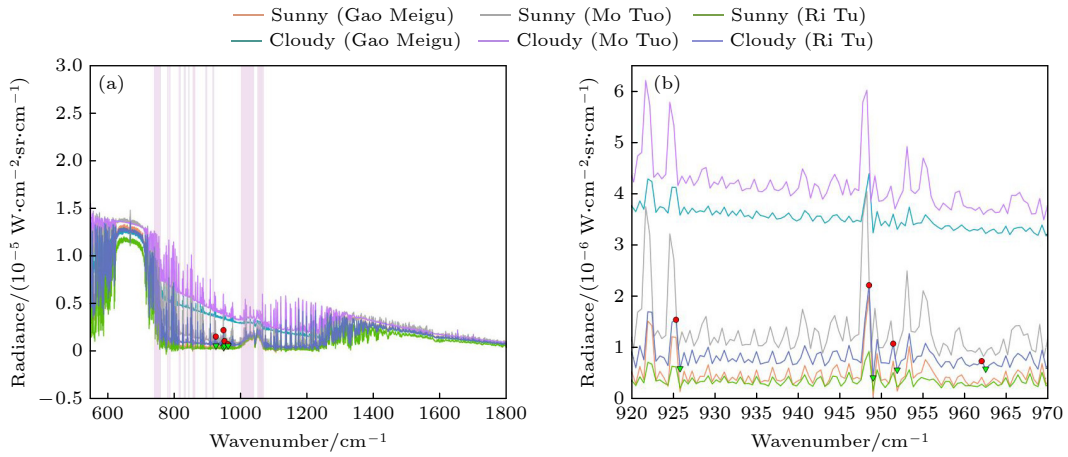

图 3 三个地点晴空和多云光谱特征 (a) 540—1800 cm^{-1} 波段晴空和多云光谱特征; (b) 920—970 cm^{-1} 波段新增的晴空和多云光谱特征的局部放大图

Fig. 3. The spectral characteristics of clear sky and cloudy conditions in three locations: (a) The spectral characteristics of clear sky and cloudy conditions in the 540–1800 cm^{-1} band; (b) a local magnified view of the newly added spectral characteristics of clear sky and cloudy conditions in the 920–970 cm^{-1} band.

表 2 用于区分多云和晴空数据的 20 个选定特征 (前 12 个特征代表原始特征, 后 8 个特征代表新增特征)

Table 2. Twenty selected features used to distinguish between cloudy and clear-sky data (the first 12 features represent the original features and the last 8 features represent the added features).

编号	特征
1	740—760 cm^{-1}波段辐亮度的斜率
2	740—760 cm^{-1}波段辐亮度的截距
3	780—920 cm^{-1}波段辐亮度的斜率
4	780—920 cm^{-1}波段辐亮度的截距
5	1000—1040 cm^{-1}波段辐亮度斜率
6	1000—1040 cm^{-1}波段辐亮度截距
7	1050—1070 cm^{-1}波段辐亮度斜率
8	784.6 cm^{-1}通道辐射与781.7—782.6 cm^{-1}波段平均辐射之间的比率
9	791.8 cm^{-1}通道辐射与789.4—790.4 cm^{-1}波段平均辐射之间的比率
10	1175 cm^{-1}和1170 cm^{-1}通道辐射之间的比率
11	1187 cm^{-1}和1184 cm^{-1}通道辐射之间的比率
12	1198 cm^{-1}和1195 cm^{-1}通道辐射之间的比率
13	925.8524 cm^{-1}通道辐亮度
14	948.9987 cm^{-1}通道辐亮度
15	951.892 cm^{-1}通道辐亮度
16	962.5007 cm^{-1}通道辐亮度
17	925.8524 cm^{-1} 和 925.3702 cm^{-1} 通道辐射之间的比率
18	948.9987 cm^{-1} 和948.5165 cm^{-1}通道辐射之间的比率
19	951.892 cm^{-1}和951.4098 cm^{-1}通道辐射之间的比率
20	962.5007 cm^{-1}和962.0185 cm^{-1}通道辐射之间的比率

3.2.1 训练集和测试集

分别随机选取三个地点晴空和多云样本的 70% 作为训练集, 30% 为测试集, 得到共 11894 个训练集样本, 5109 个测试集样本, 结果如表 3 所列.

表 3 云检测使用的训练集和测试集样本数

Table 3. The number of samples in the training set and test set used for cloud detection.

数据集	晴天样本数	多云样本数	总计
训练集(70%)	6290	5604	11894
测试集(30%)	2703	2406	5109

3.2.2 特征重要性排序

随机森林 (RF) 能在无需线性假设的前提下, 稳健高效地评估特征对模型预测的重要性, 提升解释性与特征选择效果. 为了解单个特征如何影响云检测的准确性, 在对特征进行标准化 (即采用均值为 0、标准差为 1 的分布) 后, 使用 RF 算法来计算训练集中每个特征的重要性, 以获得云检测的不同输入的权重.

3.2.3 最优云检测模型选择与性能评估

根据特征重要性排序结果, 依次选择排序后的特征作为 SVM 算法的输入, 以激光雷达的云检测结果作为参考值. 模型中的核参数 g 和惩罚因子 C 通过使用网格搜索方法 [30] 获得, 核函数使用径向基函数 (RBF). C 和 g 的范围设置为 2^{-8}—2^{8}, 搜索步长设置为 $2^{0.8}$ [25]. 网格搜索中使用的交叉验证 (CV) 折数设置为 5. 根据搜索到的 C 和 g 值建立

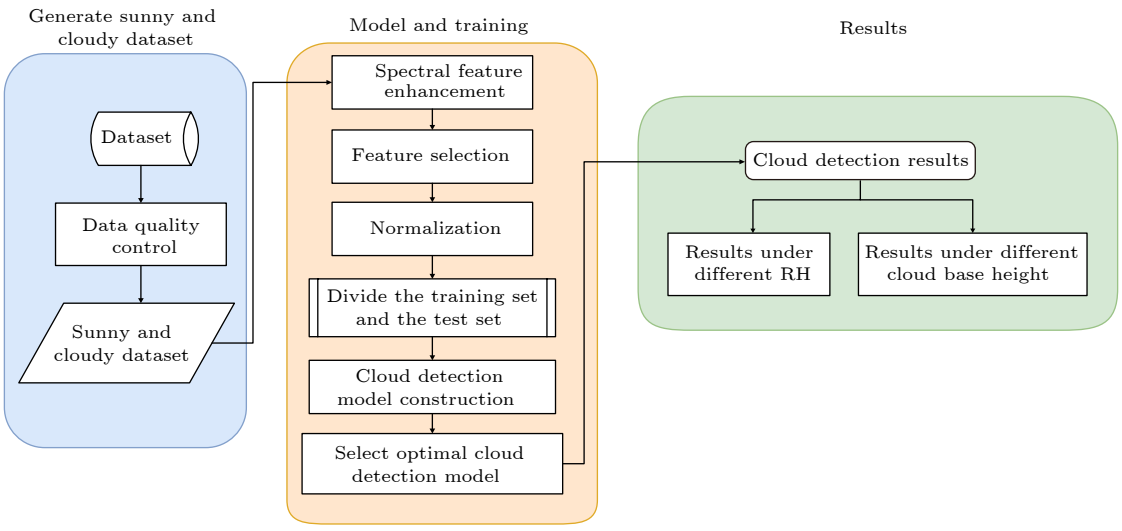

图 4 云检测算法的流程图

Fig. 4. Flowchart of the cloud detection algorithm.

云检测模型获得测试集分类的结果. 最后, 根据特征重要性排序后的不同特征构建相应的云检测模型, 使用 2×2 混淆矩阵[23] (见表 4) 进行性能评估从而得到最优云检测模型. 基于混淆矩阵, 可以得到模型分类的正确百分比 (PC)、真正率 (TPR)、真负率 (TNR)、误报率 (FPR) 以及漏报率 (FNR), 计算公式如下所示:

$$\mathrm{PC}=(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN})\times100\%, \quad (1)$$

$$\mathrm{TPR}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\times100\%, \quad (2)$$

$$\mathrm{TNR}=\mathrm{TN}/(\mathrm{FP}+\mathrm{TN})\times100\%, \quad (3)$$

$$\mathrm{FPR}=FP/(\mathrm{FP}+TN)\times100\%, \quad (4)$$

$$\mathrm{FNR}=\mathrm{FN}/(\mathrm{FN}+\mathrm{TP})\times100\%, \quad (5)$$

其中 PC 描述了所提出的算法和激光雷达检测结果之间的一致程度; TPR 可以给出 ASSIST 探测到的有云数据与激光雷达探测到的有云数据的比例, 用于描述所提出的算法对有云数据检测的灵敏度; 同样地, TNR 可以给出 ASSIST 探测到的晴空数据与激光雷达探测到的晴空数据的比例, 用于描述所提出的算法对晴空数据检测的灵敏度; TP 代表激光雷达显示有云和 ASSIST 显示有云的次数, FP 代表激光雷达显示晴空但 ASSIST 显示有云的次数, FN 代表激光雷达显示有云但 ASSIST 显示晴空的次数, TN 代表激光雷达显示晴空和 ASSIST 显示晴空的次数. FPR 代表将晴空误判为云的比例, FNR 代表将云误判为晴空的比例.

表 4 提出的算法和激光雷达云检测结果的混淆矩阵

Table 4. The confusion matrix of the proposed algorithm and lidar cloud detection results.

		激光雷达探测	
		有云	晴空
云检测算法 (ASSIST)	有云	TP (True positive)	FP (False positive)
	晴空	FN (False negative)	TN (True negative)

4 结果与讨论

4.1 模型总体性能评估

为了说明新增特征对云检测的有效性以及在不同地点的适用性, 对比分析了使用原始特征和新增特征模型在 3 个地点云检测的精度. 图 5 展示了使用原始特征和新增特征后的特征重要性排序结果, 可以看出使用原始特征和新增特征的结果完全不同, 新增的特征所占权重大于原有特征的权重. 这说明了加入新特征的重要性.

使用原始特征和新增特征排序后不同的特征个数构建不同云检测模型, 比较其云检测精度来选择最佳的云检测模型. 表 5 显示出: 对于使用原始特征构建的云检测模型, 使用前 5 个排序后的特征构建的模型云检测精度最高, PC, TPR, TNR 分别达到 95.56%, 97.26%, 94.04%. 表 6 显示出: 对于使用新增特征构建的云检测模型, 使用前 10 个排序后的特征构建的模型云检测精度最高, PC, TPR, TNR 分别达到 97.61%, 98.21%, 97.08%. 此时, 最优云检测模型的 C, g 值分别为 147.03, 9.19. 加入新的特征, 模型总体精度和晴空、多云光谱识

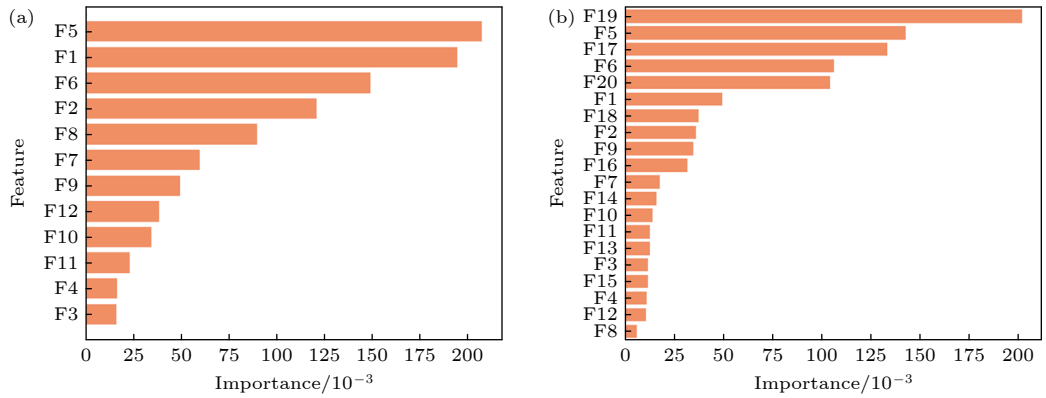

图 5 表 2 中用于区分有云和无云晴空场景的不同特征重要性排序结果 (a) 使用原始 12 个特征重要性排序结果; (b) 新增 8 个特征后重要性排序结果

Fig. 5. Importance ranking results of different features used to distinguish between cloudy and cloud-free clear sky scenes in Table 2: (a) The importance ranking results using the original 12 features; (b) the importance ranking results after adding 8 new features.

别的精度相较于使用原有特征都有所提高, 尤其是晴空光谱识别的精度有较大提升. 为了方便后续比较, 将使用原始特征选择出的最优云检测模型称为原始方法, 将新增特征选择出的最优云检测模型称为新方法.

表 5 使用原始特征排序后不同特征个数对应的云检测结果

Table 5. Cloud detection results with different numbers of features after sorting the original features.

特征个数	PC/%	TPR/%	TNR/%
1	95.01	90.90	98.67
2	95.49	93.81	97.37
3	92.88	95.84	90.23
4	94.85	96.97	92.97
5	95.56	97.26	94.04
6	85.71	97.17	75.51
7	79.43	97.22	63.60
8	86.36	97.63	76.32
9	76.59	97.67	57.82
10	76.57	97.67	57.79
11	78.74	97.88	61.71
12	81.64	98.09	67.00

表 6 使用新增特征排序后不同特征个数对应的云检测结果

Table 6. Cloud detection results with different numbers of features after sorting the newly added features.

特征个数	PC/%	TPR/%	TNR/%
1	95.30	91.98	98.26
2	94.73	94.43	95.01
3	94.72	94.43	94.97
4	96.50	94.72	98.08
5	96.24	94.80	97.52
6	96.20	96.30	96.12
7	96.46	96.76	96.19
8	96.54	97.09	96.04
9	96.56	98.09	95.19
10	97.61	98.21	97.08
11	82.60	97.38	69.44
12	95.13	97.76	92.79
13	96.81	97.42	96.26
14	96.81	97.42	96.26
15	96.59	97.42	95.86
16	88.88	97.96	80.80
17	88.49	97.88	80.13
18	80.86	98.21	65.41
19	91.41	97.76	85.76
20	91.41	97.80	85.72

4.2 不同相对湿度下性能评估

基于上述选择的最优云检测模型, 将 RH 划分为表 7 所列的 4 个不同范围, 以研究在不同水汽下云检测的精度. 首先, 根据同步观测的气象站观测数据, 将测试集在不同水汽下对应的晴空和多云样本进行划分, 详细的结果见表 7. 接着, 统计了测试集中三个地点数据在不同水汽下出现的概率. 从图 6 可以看出, 墨脱观测数据在 RH > 70% 条件下出现的概率最大, 而日土观测数据在 RH ⩽ 30% 条件下出现的概率最大, 这是因为墨脱海拔落差极大、地形复杂, 雨量极为充沛, 是中国降水最丰富的地区之一; 而日土气候严酷寒冷、降水极少, 是典型的西藏自治区高原极干冷区之一. 丽江高美古春季空气干燥, 气候宜人, 因此其观测数据也是在 RH ⩽ 30% 条件下出现的概率最大.

对划分不同 RH 范围后的测试集, 分别使用原始方法和新方法进行云检测, 得到的结果如表 8 所列. 使用新增特征构建的模型, 不管是在 RH 高

表 7 不同 RH 下测试集中三个地点总的晴空和多云样本数 (括号中的百分比表示测试集中所选 RH 范围内的数据与总测试集数据之间的比例)

Table 7. The total number of clear-sky and cloudy samples at the three locations in the test set under different RH conditions (The percentages in parentheses indicate the proportion of data within the selected RH range in the test set to the total test set data).

不同水汽	测试集 晴空样本	测试集 多云样本	总计
RH ⩽ 30%	1883	1060	2943(57.6%)
30% < RH ⩽ 50%	250	79	329(6.4%)
50% < RH ⩽ 70%	327	158	485(9.5%)
RH > 70%	243	1109	1352(26.5%)

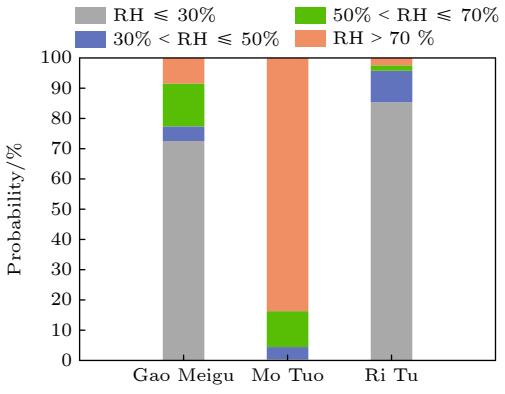

图 6 测试集中三个地点数据在不同 RH 条件下出现的概率

Fig. 6. The probability of the data from the three locations in the test set appearing under different RH conditions.

表 8 ASSIST 和激光雷达在不同 RH 条件下云检测结果的一致性

Table 8. Consistency of cloud detection results by ASSIST and lidar under different RH conditions.

不同RH	方法	PC/%	TPR/%	TNR/%	FPR/%	FNR/%
RH ⩽ 30%	原始方法	94.33	94.53	94.21	5.79	5.47
	新方法	97.93	96.89	98.51	1.49	3.11
30% < RH ⩽ 50%	原始方法	94.53	93.67	94.80	5.20	6.33
	新方法	96.66	94.94	97.20	2.80	5.06
50% < RH ⩽ 70%	原始方法	98.76	99.37	98.47	1.53	0.63
	新方法	99.58	99.40	100.00	0	0.60
RH > 70%	原始方法	97.34	99.82	86.01	13.99	0.18
	新方法	98.82	99.83	91.89	8.11	0.17

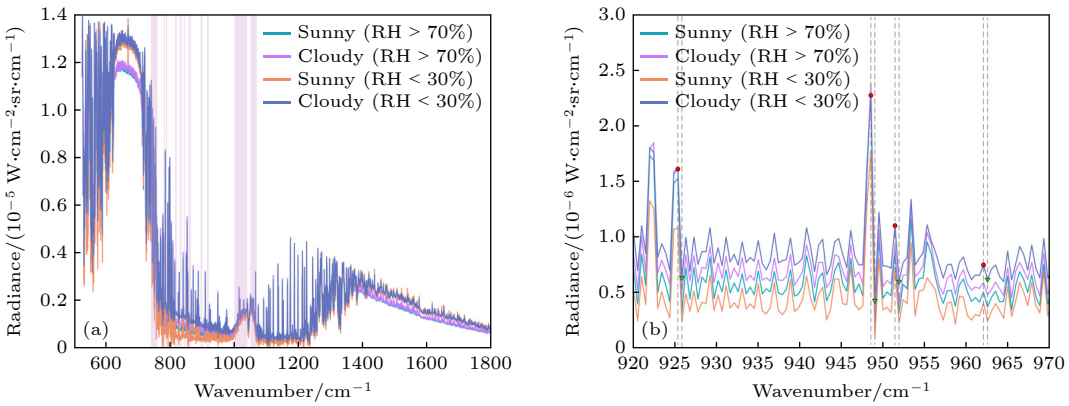

图 7 不同 RH 下, ASSIST 在高美古所测晴空和多云光谱的特征 (a) 540—1800 cm^{-1} 波段晴空和多云光谱特征; (b) 920—970 cm^{-1} 波段新增的晴空和多云光谱特征的局部放大图

Fig. 7. The spectral characteristics of clear sky and cloudy sky measured by ASSIST in Gao Meigu under different RH conditions: (a) The spectral characteristics of clear sky and cloudy conditions in the 540–1800 cm^{-1} band; (b) a local magnified view of the newly added spectral characteristics of clear sky and cloudy conditions in the 920–970 cm^{-1} band.

还是低的情况下, 晴空和多云光谱识别的精度都比原始方法识别的精度高, 尤其是对晴空光谱识别的精度有较大提高. 这是因为在晴空下不同 RH 的差异会造成晴空光谱误判. 从图 7 展示的结果可以看出: 当 RH > 70% 时, ASSIST 所测的晴空光谱特征和多云光谱特征相似, 加入新特征, 即使在 RH 较高时, 新方法也能精确识别出晴空光谱, TNR 从 86.01% 提高到了 91.89%, FPR 从 13.99% 降低到了 8.11%.

4.3 不同云底高度下性能评估

ASSIST 在高云条件下测得的光谱特征和在晴空条件下的光谱特征相似 (见图 8), 增加了云检测的难度. 为了检验加入新特征的有效性, 将同步观测的激光雷达提供的 CBH 数据划分成 4 个范围 (见表 9) 进行分析. 同样地, 我们也统计了测试集中三个地点数据在不同 CBH 下出现的概率. 图 9 结果显示: 测试集中高美古观测样本出现 1 km < CBH ⩽ 3 km 的概率最大, 而墨脱、日土观测样本出现 CBH ⩽ 1 km 的低云概率最大. 此外, 观测期间三个地方的观测数据出现 CBH > 5 km 的高云概率较小, 因此用于测试的高云样本较少.

表 9 不同 CBH 下测试集中 3 个地点总的多云样本数 (括号中的百分比表示测试集中所选 CBH 范围内的数据与总测试集数据之间的比例)

Table 9. The total number of cloudy samples at the three locations in the test set under different cloud base height conditions (The percentages in parentheses indicate the proportion of data within the selected cloud base height range in the test set to the total test set data).

不同CBH	测试集多云样本
CBH ⩽ 1 km	1196(49.69%)
1 km < CBH ⩽ 3 km	494(20.52%)
3 km < CBH ⩽ 5 km	646(26.86%)
CBH > 5 km	70(2.93%)

基于划分的不同 CBH 范围的测试集, 开展云检测, 得到的结果如表 10 所列. 结果表明: 使用新增特征构建的模型, 不管是在 CBH 高还是低的情况下, 云检测精度都高于原方法的精度, 尤其是对 3 km < CBH ⩽ 5 km 的中云和 CBH > 5 km 的高云

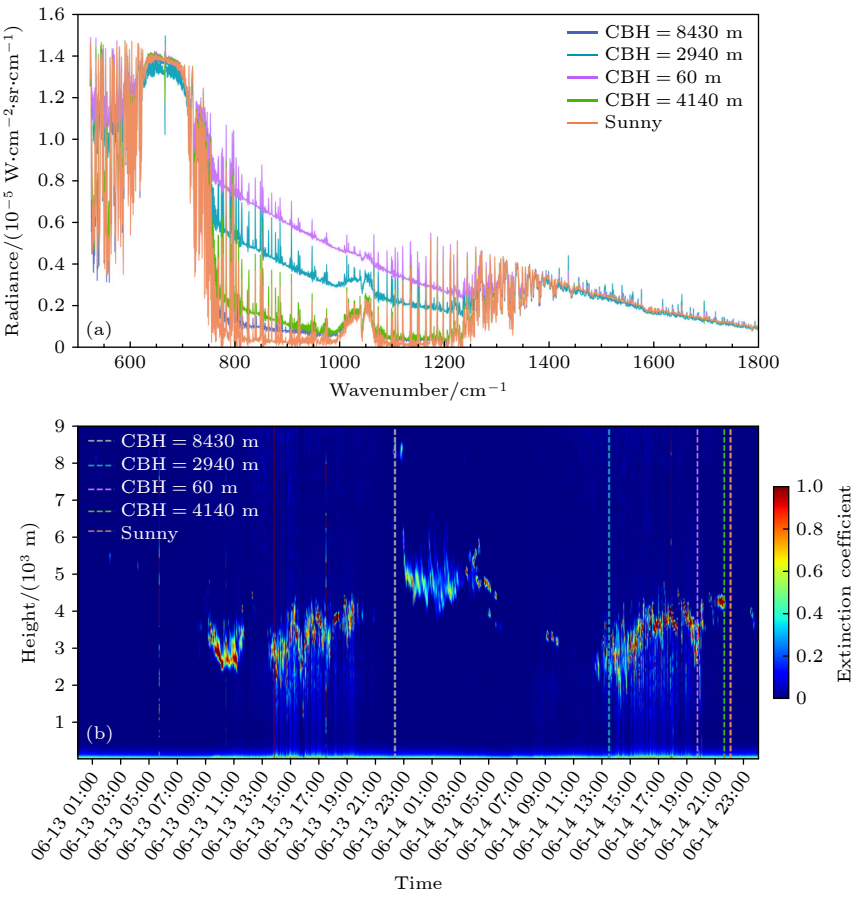

图 8 ASSIST 在日土测得的晴空和不同 CBH 下多云光谱以及激光雷达探测的云和气溶胶总消光系数结果 (a) ASSIST 测得的晴空和多云光谱; (b) 激光雷达探测的云和气溶胶总消光系数结果

Fig. 8. The results of the ASSIST measurements of clear sky and cloudy spectra, as well as the total extinction coefficients of clouds and aerosols detected by the lidar under different CBH conditions: (a) The clear sky and cloudy spectra measured by ASSIST; (b) the total extinction coefficient results of clouds and aerosols detected by lidar.

光谱识别的精度有较大提高, 即使在样本数量较少的情况下. 当 3 km < CBH ⩽ 5 km, PC 从 95.45% 提高到 98.64%, FNR 从 4.55% 降低到 1.36%; 当 CBH > 5 km, PC 从 87.5% 提高到 91.67%, FNR 从 12.5% 降低到 8.33%.

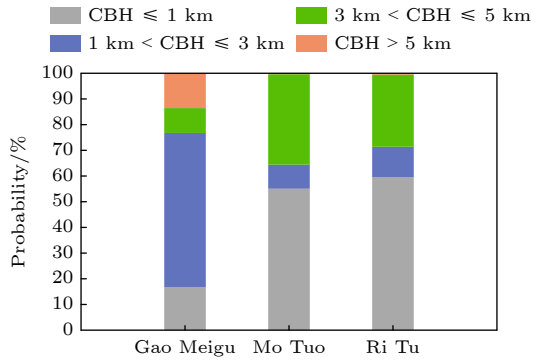

图 9 测试集中 3 个地点数据在不同 CBH 条件下出现的概率

Fig. 9. The probability of the data from the three locations in the test set appearing under different cloud base height conditions.

表 10 ASSIST 和激光雷达在不同 CBH 条件下云检测结果的一致性

Table 10. Consistency of cloud detection results by ASSIST and lidar under different cloud base height conditions.

不同CBH	方法	PC/%	TPR/%	FNR/%
CBH ⩽ 1 km	原始方法	98.53	98.53	1.47
	新方法	99.26	99.26	0.74
1 km < CBH ⩽ 3 km	原始方法	96.43	96.43	3.57
	新方法	97.62	97.62	2.38
3 km < CBH ⩽ 5 km	原始方法	95.45	95.45	4.55
	新方法	98.64	98.64	1.36
CBH > 5 km	原始方法	87.50	87.50	12.5
	新方法	91.67	91.67	8.33

4.4 案例研究

尽管新方法的云检测精度较原方法有所提升, 但其仍可能将部分无云的场景 (如雾、飘雪等) 误判为有云. 基于观测期间的实际天气状况, 本研究将重点分析有雾场景对云检测算法性能的影响.

4.4.1 薄雾

图 10 展示了 2025 年 3 月 25 日 ASSIST 在墨脱测得的晴空和薄雾光谱的案例, 激光雷达测得的消光系数在相应的时刻相对较高. 图 11 中全天空成像仪拍摄的结果也显示出那段时间是无云的. 在薄雾情况下, ASSIST 测得的光谱特征和晴空光谱非常相似, 我们提出的方法并没有将薄雾光谱判别为有云光谱, 这说明薄雾对该方法的云检测精度几乎没有影响.

4.4.2 厚雾

图 12 展示了 2025 年 3 月 24 日 ASSIST 在墨脱测得的晴空和厚雾光谱的案例. 激光雷达在相应时刻探测到的消光系数出现异常高值. 图 13 中全天空成像仪拍摄到的图像结果也显示出那段时间存在厚雾. 在厚雾情况下, ASSIST 测得的光谱特征和厚云光谱非常相似, 导致本文方法将厚雾光谱误判为有云光谱. 这是由于浓雾具有极高的粒子数密度, 显著地增强了大气下行红外辐射, 使其光谱

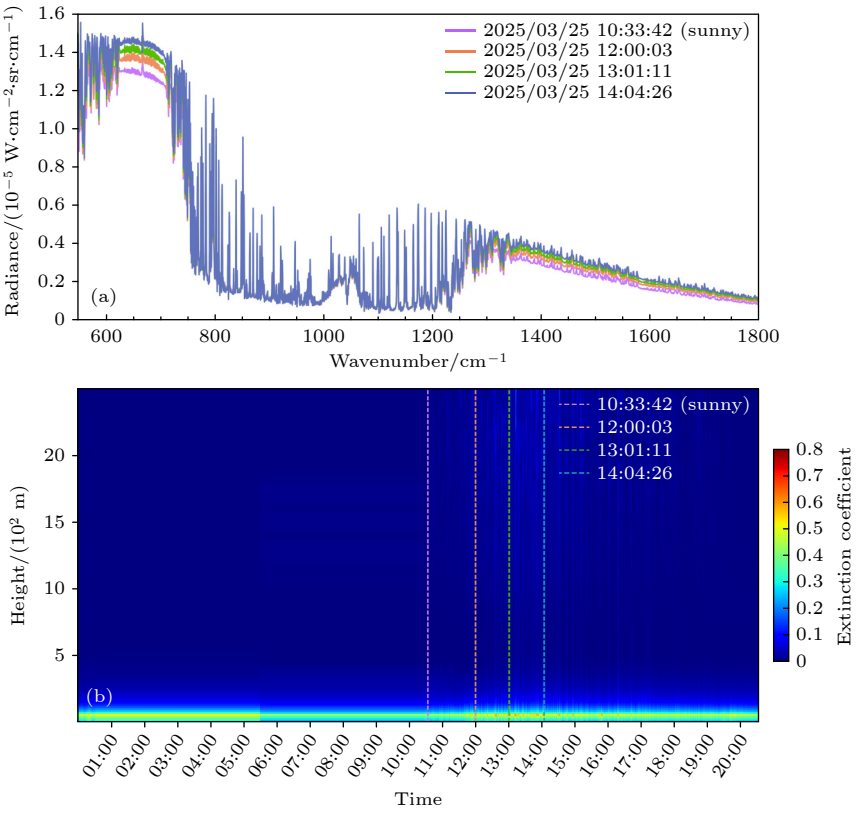

图 10 在 2025 年 3 月 25 日不同时刻, ASSIST 在墨脱测得的晴空和薄雾光谱 (a) ASSIST 测得的光谱; (b) 激光雷达探测的云和气溶胶总消光系数结果

Fig. 10. The spectra of clear sky and mist measured by ASSIST at different times in Motuo on March 25, 2025: (a) The spectra measured by ASSIST; (b) the total extinction coefficient results of clouds and aerosols detected by lidar.

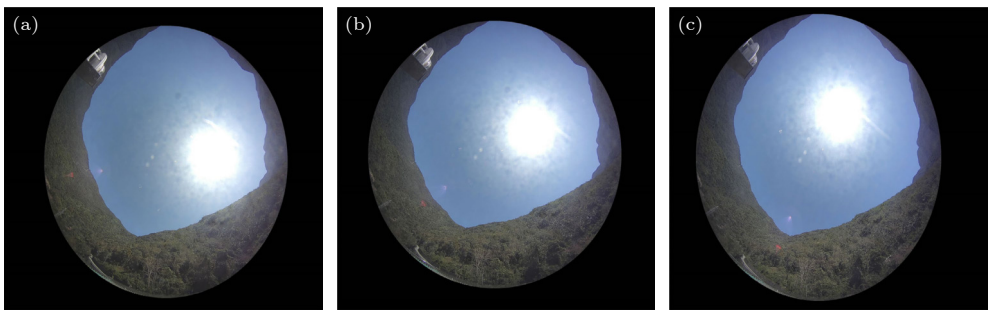

图 11 在 2025 年 3 月 25 日 3 个薄雾时刻, 全天空成像仪拍摄的图像 (a), (b), (c) 分别代表北京时间 12:00:03, 13:01:11, 14:04:26

Fig. 11. The images captured by the all-sky imager at three mist moments on March 25, 2025: (a), (b), (c) Represent Beijing time 12:00:03, 13:01:11, 14:04:26, respectively.

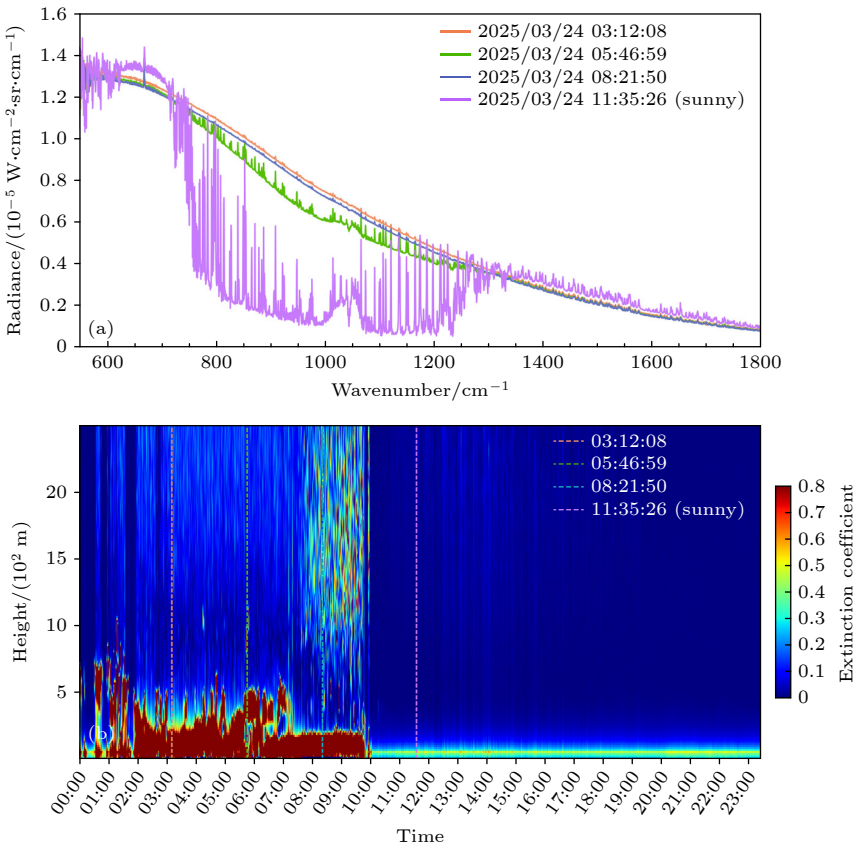

图 12 在 2025 年 3 月 24 日不同时刻, ASSIST 在墨脱测得的晴空和厚雾光谱 (a) ASSIST 测得的光谱; (b) 激光雷达探测的云和气溶胶总消光系数结果

Fig. 12. The spectra of clear sky and thick fog measured by ASSIST at different times in Motuo on March 24, 2025: (a) The spectra measured by ASSIST; (b) the total extinction coefficient results of clouds and aerosols detected by lidar.

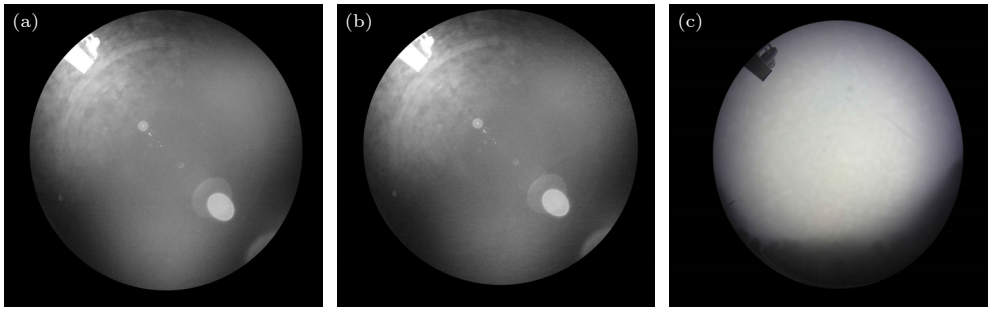

图 13 在 2025 年 3 月 24 号 3 个厚雾时刻, 全天空成像仪拍摄的图像 (a), (b), (c) 分别代表北京时间 03:12:08, 05:46:59, 08:21:50

Fig. 13. The images captured by the all-sky imager at three thick fog moments on March 24, 2025: (a), (b), (c) Represent Beijing time 03:12:08, 05:46:59, 08:21:50, respectively.

特征趋近于云的光谱特征. 因此, 厚雾对本文的云检测方法的精度构成显著影响.

5 结 论

针对地基红外高光谱云检测易受水汽干扰、高云检测精度低的问题, 本文利用 ASSIST 在丽江高美古天文台、墨脱气象观测站和日土县阿里荒漠环境综合观测站获取的观测数据, 深入分析了区分有云和晴空场景的光谱特征, 提出了一种光谱特征增强驱动的机器学习云检测方法. 通过研究该方法在不同 RH、不同 CBH 条件下与激光雷达同步探测结果的一致性, 验证了其有效性. 结果表明: 引入新增的光谱特征后, 云检测性能显著提升, PC, TPR, TNR 分别达到 97.61%, 98.21%, 97.08%. 在不同 RH 条件下, 新方法精度均优于原始方法, 尤其

是在 RH > 70% 时, 对晴空光谱识别的精度提升明显, TNR 从 86.01% 提高到了 91.89%. 同样, 在不同 CBH 条件下, 新方法精度也高于原始方法, 特别是对 3 km < CBH ⩽ 5 km 的中云和 CBH > 5 km 的高云光谱识别的精度改善明显, 即使样本量较少也能保持良好性能. 当 3 km < CBH ⩽ 5 km, PC 从 95.45% 提高到 98.64%; 当 CBH > 5 km, PC 从 87.5% 提高到 91.67%. 此外, 该方法能够有效区分薄雾条件下的多云和无云状态, 但在浓雾情况下无法实现云检测.

本研究主要基于 ASSIST 在高海拔、高原地区的观测数据, 未来工作将纳入更多不同地理位置和大气条件下的观测样本, 以进一步提升算法的普适性. 该方法可为地基红外高光谱辐射数据在辐射传输模拟、遥感反演和数值天气预报 (NWP) 模型同化等后续应用提供更可靠的数据基础.

参考文献

[1] Govender M, Chetty K, Bulcock H 2007 *Water S. A.* **33** 145
[2] Vorovencii I 2010 *Bull. Transilvania Univ. Bras. II: For. Wood Ind. Agric. Food Eng.* **2** 51
[3] Shimoda H, Ogawa T 2000 *Adv. Space Res.* **25** 937
[4] Aumann H H, Miller C R 2002 *Proc. SPIE* **4483** 332
[5] Clerbaux C, Hadji-Lazaro J, Turquety S, George M, Coheur P F, Hurtmans D, Wespes C, Herbin H, Blumstein D, Tourniers B, Phulpin T 2007 *Space Res. Today* **168** 19
[6] Andrew S, Nigel A, William B, Amy D 2015 *Atmos. Sci. Lett.* **16** 260
[7] Qi C L, Wu C Q, Hu X C, Xu H L, Lee L, Zhou F, Gu M J, Yang T H, Shao C Y, Yang Z D, Zhang P 2020 *IEEE Trans. Geosci. Remote Sens.* **58** 4335
[8] Yang J, Zhang Z Q, Wei C Y, Lu F, Guo Q 2017 *Bull. Am. Meteorol. Soc.* **98** 1637
[9] Knuteson R O, Revercomb H E, Best F A, Ciganovich N C, Dedecker R G, Dirkx T P, Ellington S C, Feltz W F, Garcia R K, Howell H B, Smith W L, Short J F, Tobin D C 2004 *J. Atmos. Oceanic Technol.* **21** 1763
[10] Rochette L, Smith W, Howard M, Bratcher T 2009 *SPIE* **7457** 002
[11] Turner D D, Löhnert U 2014 *J. Appl. Meteorol. Climatol.* **53** 752
[12] Mariani Z, Strong K, Palm M, Lindenmaier R, Adams C, Zhao X, Savastiouk V, McElroy C T, Goutail F, Drummond J R 2013 *Atmos. Meas. Tech.* **6** 1549
[13] Seo J, Choi H, Oh Y 2022 *Remote Sens.* **14** 407
[14] Ye J, Liu L, Yang W Y, Ren H 2022 *IEEE Geosci. Remote Sens. Lett.* **19** 1
[15] Wang Y, Xiong W, Ye H H, Shi H L, Wang X H, Li C, Wu S C, Cheng C 2025 *Remote Sens.* **17** 1440
[16] Wang Y, Ye H H, Shi H L, Wang X H, Li C, Sun E C, An Y, Wu S C, Xiong W 2024 *J. Quant. Spectrosc. Radiat. Transfer* **326** 109118
[17] Mcnally A, Watts P A 2003 *Q. J. R. Meteorol. Soc.* **129** 3411
[18] Bauer P, Auligné T, Bell W, Geer A, Guidard V, Heilliette S, Kazumori M, Kim M J, Liu E, McNally A, Macpherson B, Okamoto K, Renshaw R, Riishøjgaard L P 2011 *Q. J. R. Meteorol. Soc.* **137** 1934
[19] Löhnert U, Turner D, Crewell S 2009 *J. Appl. Meteorol. Climatol.* **48** 1017
[20] Li C, Ma J J, Yang P, Li Z Q 2019 *J. Quant. Spectrosc. Radiat. Transfer* **222** 196
[21] Cho J S, Goo T Y, Shin J 2015 *Korean Soc. Remote Sens.* **31** 137
[22] Zhang Q, Yu Y, Zhang W M, Luo T L, Wang X 2019 *Remote Sens.* **11** 3035
[23] Luo T L, Zhang W M, Yu Y, Feng M, Duan B H, Xing D 2019 *Int. J. Remote Sens.* **40** 6530
[24] Shi H X, Yu Y, Zhang W M, Ma G, Zhang Q, Luo T L, Huang Q B 2021 *Proceedings of the 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)* Kunming, China, September 17–19, 2021 p107
[25] Liu L, Ye J, Li S L, Hu S, Wang Q 2022 *Remote Sens.* **14** 2589
[26] Michaud-Belleau V, Gaudreau M, Lacoursière J, Boisvert É, Ravelomanantsoa L, Turner D, Rochette L 2025 EGUsphere: 2024-3617 [atmospheric sciences]
[27] Zhao Q, Su H C, Yi M J, Yu D S, Xu C D 2021 *Chin. J. Lasers* **48** 2010001 (in Chinese) [赵强, 苏红超, 易明建, 余东升, 徐赤东 2021 中国激光 **48** 201001]
[28] TURNER D 2007 *J. Geophys. Res. Atmos.* **112** D15
[29] Ishida H, Oishi Y, Morita K, Moriwaki K, Nakajima T 2018 *Remote Sens. Environ.* **205** 390
[30] Wang X P, Zhang F, Kung H T, Johoson V C, Latif A 2020 *Int. J. Remote Sens.* **41** 953

A spectral feature enhancement-driven machine learning method for cloud detection using ground-based infrared hyperspectral data*

WANG Yue [1)2)3)] YE Hanhan [2)3)†] XIONG Wei [1)2)3)‡] WANG Xianhua [2)3)]
SHI Hailiang [2)3)] LI Chao [4)] CHENG Chen [2)3)] WU Shichao [2)3)]

1) (*School of Environmental Science and Optoelectronic Technology, University of Science and Technology of China, Hefei* 230026, *China*)

2) (*Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei* 230031, *China*)

3) (*Anhui Province Key Laboratory of Optical Quantitative Remote Sensing, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei* 230031, *China*)

4) (*School of Earth and Space Science, University of Science and Technology of China, Hefei* 230026, *China*)

(Received 23 July 2025; revised manuscript received 11 August 2025)

Abstract

Clouds exert a significant influence on infrared radiation, making cloud detection a crucial step in the application of infrared hyperspectral data. In particular, water vapor interference and the limited accuracy in high-cloud identification constitute two key challenges for ground-based infrared hyperspectral cloud detection. Traditional threshold-based cloud detection methods are difficult to adapt to different locations and dynamically changing atmospheric conditions,while machine learning methods can achieve cloud detection with higher accuracy, greater robustness, and improved automation. Building on the advantages of machine learning, observational data from the atmospheric sounder spectrometer by infrared spectral technology (ASSIST), collected at Lijiang (Yunnan), Motuo (Xizang Autonomous Region), and Ritu (Xizang Autonomous Region) in China, are used to analyze the spectral differences between sunny and cloudy conditions in this study. Based on these differences, a spectral feature enhancement-driven machine learning method for cloud detection is proposed. Finally, by incorporating synchronous observations from lidar, meteorological stations, and all-sky imagers, the proposed method is systematically evaluated under different relative humidity (RH) and cloud base height (CBH) conditions. The experimental results show that the consistency between the results obtained by the proposed method and lidar-based detection is as high as 97.61%. Under different RH conditions, the proposed method outperforms the method based on original spectral features. Notably, when $\mathrm{RH} > 70\%$, the accuracy of clear-sky spectral identification improves significantly: increasing from 86.01% to 91.89%. Similarly, under different CBH conditions, the proposed method also exhibits superior performance compared with the method in which original spectral features are used. In particular, the accuracy improvements are especially notable when identifying mid-level clouds with $3\ \mathrm{km} < \mathrm{CBH} \leqslant 5\ \mathrm{km}$, as well as high-level clouds with $\mathrm{CBH} > 5\ \mathrm{km}$. When $3\ \mathrm{km} < \mathrm{CBH} \leqslant 5\ \mathrm{km}$, the accuracy increases from 95.45% to 98.64% and when $\mathrm{CBH} > 5\ \mathrm{km}$, the accuracy improves from 87.5% to 91.67%. The proposed method significantly enhances the automation and accuracy of cloud detection, thereby providing higher-quality fundamental datasets for supporting subsequent applications such as radiative transfer simulation, remote sensing parameter retrieval, and data assimilation in numerical weather prediction (NWP) models.

Keywords: ground-based infrared hyperspectroscopy, remote sensing, machine learning, cloud detection

PACS: 02.70.Hm, 07.05.Mh, 07.57.Ty, 42.68.Ge **DOI:** 10.7498/aps.74.20250982

CSTR: 32037.14.aps.74.20250982

* Project supported by the National Key R&D Program of China (Grant No. 2022YFB3901804) and the Natural Science Foundation of Anhui Province, China (Grant No. 2408055UQ003).

† Corresponding author. E-mail: yehanhan@aiofm.ac.cn

‡ Corresponding author. E-mail: frank@aiofm.ac.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

一种光谱特征增强驱动的机器学习地基红外高光谱云检测方法

王越 叶函函 熊伟 王先华 施海亮 李超 程晨 吴时超

A spectral feature enhancement-driven machine learning method for cloud detection using ground-based infrared hyperspectral data

WANG Yue YE Hanhan XIONG Wei WANG Xianhua SHI Hailiang LI Chao CHENG Chen WU Shichao

引用信息 Citation: Acta Physica Sinica, 74, 200202 (2025) DOI: 10.7498/aps.74.20250982
CSTR: 32037.14.aps.74.20250982
在线阅读 View online: https://doi.org/10.7498/aps.74.20250982
当期内容 View table of contents: http://wulixb.iphy.ac.cn
