第 43 卷 第 6 期/2023 年 3 月/光学学报 研究论文

光学学报

基于集成学习的FY-4A云底高度反演方法

余茁夫[1]，王雅[2*]，马烁[1**]，艾未华[1]，严卫[1]

[1]国防科技大学气象海洋学院，湖南 长沙 410000；

[2]中国气象局国家卫星气象中心，北京 100081

摘要 云底高度是地气系统辐射收支以及飞行安全的重要影响因素。介绍了利用FY-4A卫星的数据产品反演云底高度的方法，设计了两种云底高度反演方案：第一种方案先将云划分为卷云(Ci)、高层云(As)、高积云(Ac)、层云/层积云(St/Sc)、积云(Cu)、雨层云(Ns)、深对流云(Dc)和多层云(Multi)等8种云类型，再分别采用独立的集成学习模型反演这8类云的云底高度；第二种方案不区分云的类型，采用统一的集成学习模型反演云底高度。将CloudSat探测的云底高度作为参考值，以129515个样本对两种方案进行评估，结果表明方案一的反演模型效果更好，均方根误差(RMSE)为1304.7 m，平均绝对误差(MAE)为898.4 m，相关系数(R)为0.9214。

关键词 大气光学；云底高度反演；FY-4A；云顶高度；云光学厚度；云粒子有效半径；集成学习

中图分类号 P412 **文献标志码** A **DOI**：10.3788/AOS220957

1 引 言

作为云的宏观物理参数之一，云底高度(CBH)对地气系统的辐射收支具有重要的调节作用，Viúdez-Mora等[1]指出，云底高度对地表接收的下行长波辐射的影响仅次于大气的温度和湿度。另外，云底高度与能见度密切相关，在飞行活动中，过低的云底高度会引起能见度下降，当能见度下降到一定程度，飞行员就会失去辨明飞行高程的视觉条件，极易引发飞行事故[2]。因此，在科学研究和相关气象业务中，获取准确的云底高度具有非常重要的意义。

CloudSat卫星上搭载的毫米波雷达——cloud profile radar(CPR)能够探测云的垂直分布信息，其云底高度的数据在科学研究和气象业务中得到了非常广泛的应用[3-4]。与CloudSat主动遥感的工作方式不同，气象卫星上搭载的探测仪器大多是被动遥感仪器，利用卫星被动遥感资料生产的云产品主要是云顶产品，很少有云底产品[5]。因此，研究如何利用卫星的被动遥感资料来反演云底高度对卫星云底产品的开发具有重要的参考意义。Hutchison[6]介绍了一种利用中分辨率成像光谱仪(MODIS)数据反演云底高度的方法。他将云分成水云和冰云两类，利用MODIS产品的云粒子半径、云光学厚度和云顶温度，结合Liou[7]总结的经验公式，得到水云和冰云的厚度，最后用云顶高度减去云层厚度得到云底高度。该算法被应用到Suomi National Polar Orbiting Partnership(SNPP)上的Visible Infrared Imager Radiometer Suite(VIIRS)上，成功开发出SNPP/VIIRS的云底高度业务产品。但是，Seaman等[8]指出SNPP/VIIRS的云底高度产品在实际应用中的效果并不理想。为此，Noh等[9]设计了一种基于统计的SNPP/VIIRS云底高度反演算法，他们对云层厚度、云顶高度和云水路径进行统计分析，建立回归方程。在统计中，他们没有区分云的类型，但是他们只将回归方程用于非卷云和非深对流的云；对于卷云和深对流云，他们采用其他方式得到云底高度。

Forsythe等[10]提出一个基本假设：同类型的云具有相近的云底高度。他们利用卫星的云类型产品得到地面站点上空的云类型，然后根据地面站点对云底高度的观测，外推卫星视场内同类云的云底高度。在Forsythe等研究的基础上，王帅辉等[11]利用MODIS的光谱数据得到云类型的分布，根据CloudSat的云类型及云底高度对MODIS视场内同类云的云底高度进行外推估计。李浩然等[12]根据云顶高度和云水路径对CloudSat和MODIS的数据进行模板匹配，并据此估计出MODIS的云层厚度，再用云顶高度减去云层厚度得到云底高度。

上述云底高度的研究都是以极轨卫星为对象，而以静止气象卫星为对象的研究还比较少。高顶[13]和谭仲辉等[14]分别对FY-4A的云底高度反演算法进行了研究。他们根据FY-4A的云相态产品和FY-4A的云

收稿日期：2022-04-13；**修回日期**：2022-06-29；**录用日期**：2022-08-01；**网络首发日期**：2022-08-10

基金项目：国家自然科学基金(41705007)

通信作者：*ywang@cma.gov.cn；**mashuo0601@163.com

类型产品划分出不同类型的云，然后分别采用独立的模型反演各类云的云底高度。但是，FY-4A的云相态产品和云类型产品无法区分出深对流云，云底高度的反演效果受到了较大的影响[14]。Tan等[15]研究了Himawari-8的云底高度反演算法：对于非深对流的云，不区分云的类型，采用统一的模型反演云底高度；对于深对流云，用抬升凝结高度代替其云底高度。他们虽然提出了针对深对流云的处理方式，但是没有对深对流云云底高度的反演效果进行评估。Lin等[16]介绍了Geostationary Operational Environmental Satellite (GOES)-16的云底高度估计方法，但在研究中也没有区分云的类型，直接利用GOES-16的一级数据采用统一的模型对云底高度进行估计。

综合之前的研究，有一个问题需要重点关注，即在反演云底高度时，是否需要区分云的类型。从上述研究的结果来看，区分云类型和不区分云类型这两种思路均能够取得较好的效果。为了得到更准确的云底高度，本文对这两种思路进行比较分析，希望从中找到更适合于FY-4A的云底高度反演思路。

2 数　　据

FY-4A是我国第二代静止气象卫星的首发星，于2016年12月11日发射，定位于104.7°E上空[17]。advanced geostationary radiation imager(AGRI)是FY-4A的主要载荷之一，共设置14个光谱通道，其范围为0.45～13.8 μm，空间分辨率为0.5～4 km，表1列出了AGRI上14个通道的相关参数。

FY-4A/AGRI大幅提升的时空分辨率极大地增强了卫星的观测能力及参数反演能力。如前文所述，之前的研究在反演云底高度时大多利用了卫星的二级产品，如云顶高度、云光学厚度和云粒子有效半径等。因此，本文也主要利用云顶高度、云光学厚度和云粒子有效半径等数据产品来研究FY-4A的云底高度反演方法。中国气象局国家卫星气象中心已经利用FY-4A/AGRI的多通道数据开发出这些产品，Xu等[18]和Lai等[19]总结了这三种产品所用的光学通道，如表2所示，这3种产品的空间分辨率均为4 km。目前，这些云产品已经应用到相关的研究中[20-25]。

作为全球首个搭载毫米波测云雷达的卫星，CloudSat自2006年发射以来就一直受到学界的高度关注。在这十几年中，CloudSat卫星的数据产品不断被验证、改进，其质量得到了广泛的认同和肯定。在大气遥感的相关研究中，CloudSat卫星的数据产品常作为验证数据来检验其他算法和产品的准确性。例如：Seaman等[8]在介绍VIIRS云底高度的业务算法和产品时，将CloudSat的云底高度作为真值来验证VIIRS的云底高度业务产品；Kahn等[4]在对atmospheric infrared sounder(AIRS)反演的云高度进行精度评估时，同样将CloudSat的数据作为参考标准。因此，本文也利用CloudSat的数据产品来研究FY-4A的云底高度反演方法，并对反演结果进行验证。前面提到，CloudSat/CPR能够获取云的垂直分布信息。CPR的每条扫描轨迹包含了约36950个像素点(每条轨迹上像素点的数量并不完全相同)，每一个像素点都对应一条垂直廓线，在垂直廓线上共有125个库，每个库的垂直分辨率为240 m。每个像素点在沿轨道方向上的水平分辨率为2.5 km，在跨轨道方向上的分辨率为1.4 km。本文使用CloudSat的2B-GEOPROF产品和2B-CLDCLASS产品来获取云底高度和云类型。

FY-4A和CloudSat在探测云层信息时各有优劣：FY-4A观测范围广，时间分辨率高，但是FY-4A/AGRI是被动遥感仪器，很难探测到云层底部的信息；CloudSat采用主动遥感的方式对大气进行观测，能够探测到云层底部的信息，由于CloudSat是极轨气象卫星，虽然能够获取全球范围内的云信息，但是它对地球上某一点进行两次观测的时间间隔较长，观测的连续性较低。因此，本文结合二者的观测优势，利用这两个卫星的相关数据开展FY-4A云底高度反演方法的

表1　FY-4A/AGRI光谱通道的参数

Table 1　Parameters of FY-4A/AGRI channels

Channel number	Center wavelength / μm	Spatial resolution / km
1	0.47	1.0
2	0.65	0.5
3	0.825	1.0
4	1.375	2.0
5	1.61	2.0
6	2.25	2.0
7	3.75 (H)	2.0
8	3.75 (L)	4.0
9	6.25	4.0
10	7.1	4.0
11	8.5	4.0
12	10.7	4.0
13	12.0	4.0
14	13.5	4.0

表2　FY-4A/AGRI的相关云产品及所用的光学通道

Table 2　Related cloud products of FY-4A/AGRI and corresponding channels

Cloud product of FY-4A/AGRI	Center wavelength of optical channel /μm
Cloud top height (CTH)	11.0, 12.09, 13.55
Cloud optical thickness (COT) and cloud effective radius (CER)	0.65, 2.25

研究。

首先，根据CloudSat/CPR的产品确定云底高度和云类型。在2B-GEOPROF产品中有一个CPR_Cloud_Mask变量，这个变量是根据CPR的反射率因子得到的云覆盖数据。通过这个变量，可以确定在每个像素点的垂直廓线上是否有云层存在。前面提到，每条垂直廓线有125个库，在每个库上都有CPR_Cloud_Mask变量的值，值的大小通常在0到40之间。该变量的值大于5，表示这个库所在的高度可能有水滴存在，值越大表示探测的准确性越高。根据CloudSat的算法文档[26]，将20设为一个临界值，当某个库上CPR_Cloud_Mask的值大于20，认为这个库所在的高度上有云层存在，这个像素点即为有云像素点。自地面向上，云层存在的第一个库所在的高度即为云底高度。

然后，判断CloudSat/CPR有云像素点的云类型。在2B-CLDCLASS产品中，云类型的信息储存在cloud_scenario变量中。在某个有云像素点的垂直廓线上，如果某种类型的云分布在连续两个或更多的库上，就将其确定为该有云像素点的云类型。如果在该有云像素点上有多种不同类型的云，就认为该有云像素点的上空是多层云。

最后，参考Min等[27]和Lai等[19]的时空匹配方法，对CloudSat/CPR的有云像素点和FY-4A/AGRI的像素点进行时空匹配，图1为数据匹配的流程图。AGRI中一个时次的云产品的时间跨度为15 min，产品中每个像素点的水平分辨率为4 km。如果AGRI像素点和CPR像素点的时间间隔在15 min以内，且空间间隔在2 km（AGRI像素点水平分辨率的半径）以内，认为两颗卫星的像素点是时空匹配的。对一个CPR像素点，

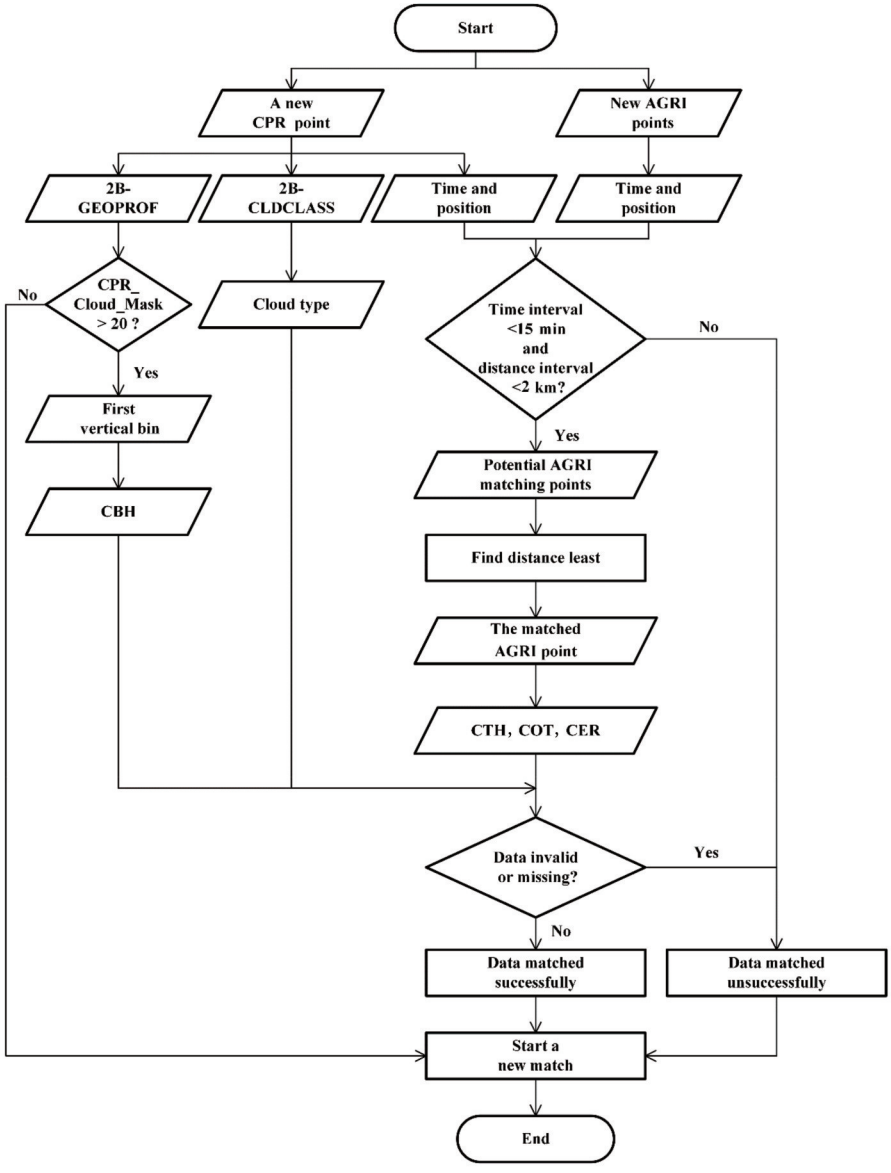

图1 FY-4A和CloudSat的数据匹配流程图

Fig. 1 Flow chart of data matching between FY-4A and CloudSat

可能存在多个满足条件的AGRI像素点，选择其中与CPR像素点距离最小的一个作为最终的匹配点。如果该CPR像素点的CBH和云类型等数据是有效值，且其匹配AGRI像素点的CTH、COT和CER等数据也是有效值，则认为这次匹配过程是成功的。本实验对2018年和2019年的数据进行匹配，最终共得到431713组样本数据。

3 方　法

3.1 方案设计

设计了FY-4A云底高度的两种反演方案：第一种方案是根据云类型反演云底高度。需要注意的是，这里的云类型不是FY-4A现有云类型产品的类型，而是CloudSat云类型产品的类型，即卷云(Ci)、高层云(As)、高积云(Ac)、层云(St)、层积云(Sc)、积云(Cu)、雨层云(Ns)、深对流云(Dc)和多层云(Multi)。对于多层云，只关注两层云的场景且主要考虑底层云的云底高度。在431713组匹配样本中，各类云的样本数量如表3所示。第二种方案是不考虑云的类型直接采用统一的模型反演云底高度。

表3　数据匹配后得到的各类云的数量

Table 3　Numbers of all types of clouds after data matching

Cloud type	Number
Ci	73255
As	71000
Ac	41412
St/Sc	104792
Cu	24535
Ns	40397
Dc	10719
Multi	65603

在CloudSat的云分类算法文档里，Wang[26]提到目前的CloudSat云分类算法不能很好地将2B-CLDCLASS产品中的St和Sc区分开，他建议最好将St和Sc结合在一起进行研究。因此，本文将St和Sc作为同一种云类型进行研究。

Hutchison[6]设计的MODIS云底高度反演算法的物理本质是用云顶高度减去云层厚度得到云底高度，而如何获取云层厚度是算法的核心问题。Hutchison根据相关经验公式，利用MODIS相关的云产品计算出云层的厚度。科研人员将这种算法应用于SNPP/VIIRS，开发出云底高度的业务产品。在该算法中，云顶高度、云粒子有效半径和云光学厚度之间是非线性相关的关系，机器学习算法能够很好地处理这种非线性相关性。因此，本文基于云顶高度、云层厚度和云底高度的几何关系这一物理本质，根据集成学习的算法原理，利用FY-4A的CTH、COT和CER探讨云底高度的反演算法。

3.2 集成学习算法

在机器学习中，集成学习算法是通过构建并结合多个学习器来完成学习任务的一种算法。由于结合了多个学习器，集成学习算法常可获得比单一学习器更加显著的泛化性能[28]。目前常用的集成学习算法大致可分为两大类：一类是个体学习器间不存在依赖关系，可同时生成的并行化方法，该类方法以bagging方法为代表；另一类是个体学习器之间存在依赖关系，必须串行生成的序列化方法，该类方法以boosting方法为代表。

bagging方法是一种有放回的随机抽样方法。通过这种抽样方法从总的样本集中抽取出多个不同的样本子集，在这些样本子集上分别训练个体学习器，然后将所有训练好的个体学习器组合成最终的模型。随机森林(RF)算法是bagging算法的典型代表，由Breiman[29]在2001年提出。RF算法以决策树为个体学习器。在决策树结点划分时，RF算法先从所有特征中随机选取若干个特征，然后根据某个指标从这若干个特征中选择最优的特征来对样本进行划分[30]。因此，RF算法在结点划分时引入了随机性，有利于算法泛化性能的提高，这也是RF得名的由来。

boosting方法在训练过程中使用的样本是固定的。在每一轮训练过程中，根据上一轮个体学习器的训练结果，调整本轮个体学习器的训练方向，并将所有个体学习器加权组合形成最终的模型。梯度提升树(GBT)算法是Freidman[31]提出的一种boosting方法，它同样以决策树为个体学习器，其原理类似于最速下降法。每轮训练过程结束后，都能得到预测值与目标值之间的偏差。在本轮训练过程中，按照上一轮偏差梯度的方向训练新的决策树，这样能够保证本轮偏差低于上一轮的偏差。将每一轮训练的决策树组合起来即得到最终的模型。

3.3 反演方案

图2给出了两种云底高度反演方案，这里主要采用RF和GBT两种集成学习模型来建立对应的云底高度反演模型。

方案一按照云类型反演云底高度。首先将样本按照云类型区分开，然后按照7∶3的比例将每类云的样本随机分为训练验证集和测试集。对于每类云的样本，首先采用十折交叉验证的方式在训练验证集上确定模型的参数。然后，根据确定的参数利用训练验证集的样本训练模型。最后，在测试集上评估两个模型的效果，将效果最好的模型作为该类云的云底高度反演模型。方案二不区分云的类型，将方案一中每类云在训练验证集上的样本合在一起作为方案二的训练验证集，将测试集的样本合在一起作为方案二的测试集。在训练验证集上，通过十折交叉验证确定模型的参数并训练最终的模型，再利用测试集评估两个模型的效果，将效果最好的模型作为方案二最终的云底高度反

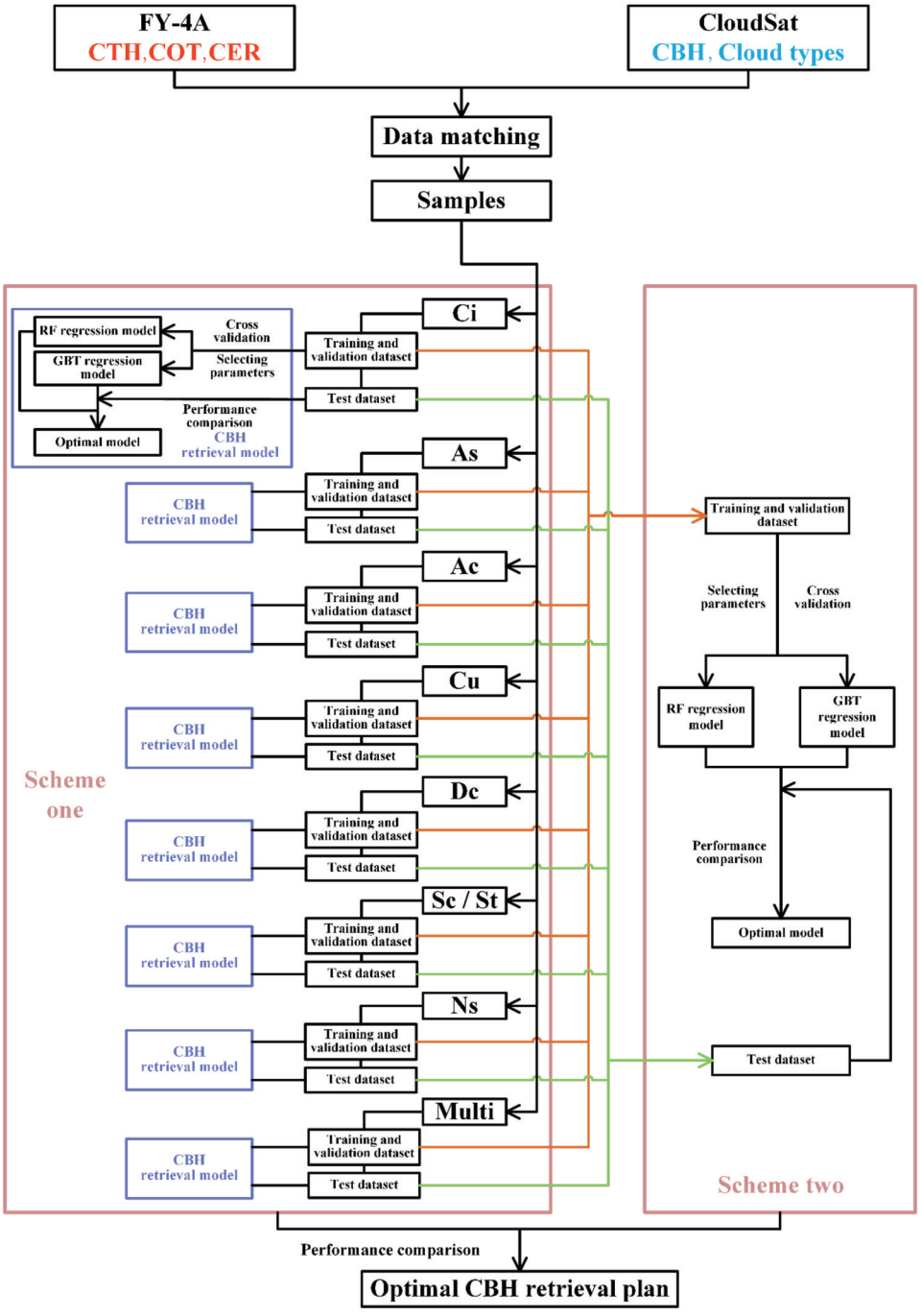

图 2　FY-4A 云底高度反演方案

Fig. 2　Schemes of CBH retrieval for FY-4A designed in this paper

演模型。根据两种方案在测试集上的效果，确定最终的云底高度反演模型。

4　结果与讨论

根据模型结果和真实值之间的均方根误差(RMSE)来确定模型参数。在评价模型和方案的效果时，还考虑了平均绝对误差(MAE)、相关系数(在后文中用 R 表示)和平均相对误差(MRE)3个指标。对RF模型和GBT模型，主要考虑的参数是决策树的数量和最大深度。

4.1　方案一

按照图2所示的反演方案，需要在方案一中对每类云分别确定RF模型和GBT模型的参数。先将决策树的最大深度设为3(该取值是随机选择的)，然后在每类云的训练验证集上分别训练这两个模型。根据训练验证集上样本的RMSE确定决策树的数量。图3所示为两个模型的决策树数量在不同取值时，训练验证集上每类云样本的RMSE变化情况。

在RF模型中，随着决策树数量的增加，各类云的样本在训练验证集上的RMSE均表现出先减小后基

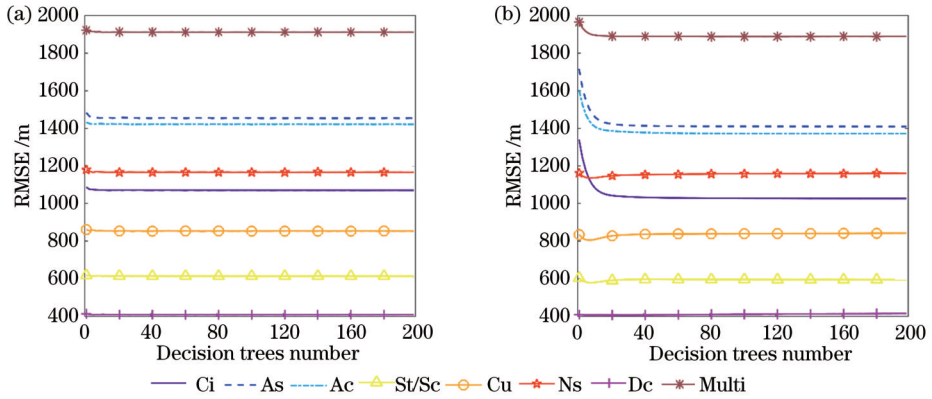

图 3　8 类云的 RMSE 随 RF 模型和 GBT 模型中决策树数量的变化。(a) RMSE 随 RF 模型中决策树数量的变化；(b) RMSE 随 GBT 模型中决策树数量的变化

Fig. 3　Variation of RMSE of eight types of clouds with decision trees number in RF model and GBT model. (a) Variation of RMSE with decision trees number in RF model; (b) variation of RMSE with decision trees number in GBT model

本稳定的变化趋势。在 GBT 模型中，随着决策树数量的增加，Ci、As、Ac 和 Multi 的 RMSE 表现出先减小后基本稳定的变化趋势，而 St/Sc、Cu、Ns 和 Dc 的 RMSE 则是先减小后增大。当 RMSE 的变化不超过 1 m 时，将对应的决策树数量确定为最佳取值。据此，对每类云分别确定了这两个模型中决策树数量的最佳取值，结果如表 4 所示。

根据表 4 的结果设置决策树的数量，并重新在各类云的训练验证集上分别训练这两个模型，分析样本的 RMSE 随决策树最大深度的变化。图 4 所示为两个模型的决策树最大深度在不同取值时，训练验证集上每类云样本的 RMSE 变化情况。

从图 4 可以看到，当决策树的最大深度从 1 增加到 20 时，各类云的 RMSE 均表现出先减小后增大的变化趋势，在决策树的某个深度时，RMSE 达到最小。在两个模型中，样本的 RMSE 均呈现出这种变化趋势。据此，对每类云分别确定这两个模型中决策树的最大深度，结果如表 5 所示。

表 4　针对每类云设置的两个模型的决策树数量

Table 4　Decision trees number of two models for all types of clouds

Cloud type	Decision trees number of RF model	Decision trees number of GBT model
Ci	13	93
As	9	71
Ac	13	83
St/Sc	11	7
Cu	13	7
Ns	13	7
Dc	5	3
Multi	11	27

至此，为各类云的两个模型确定了相应的参数。根据表 4 和表 5 所示的设置参数，在训练验证集上对各类云的两个模型进行最终训练，并利用测试集上的样

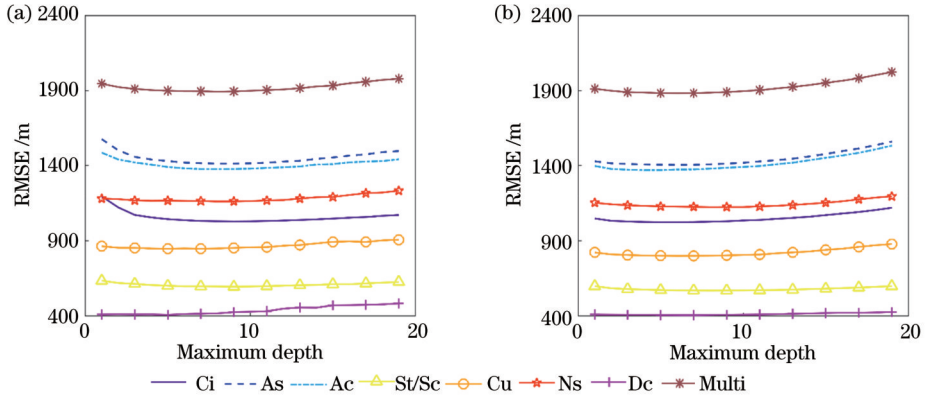

图 4　8 类云的 RMSE 随 RF 模型和 GBT 模型中决策树最大深度的变化。(a) RMSE 随 RF 模型中决策树最大深度的变化；(b) RMSE 随 GBT 模型中决策树最大深度的变化

Fig. 4　Variation of RMSE of eight types of clouds with maximum depth of decision trees in RF model and GBT model. (a) Variation of RMSE with maximum depth of decision trees in RF model; (b) variation of RMSE with maximum depth of decision trees in GBT model

表 5 针对每类云设置的两个模型的决策树最大深度

Table 5 Maximum depth of decision trees of two models for all types of clouds

Cloud type	Maximum depth in RF	Maximum depth in GBT
Ci	8	5
As	8	5
Ac	7	4
St/Sc	9	6
Cu	4	5
Ns	8	8
Dc	5	6
Multi	8	5

本选择最优的 CBH 反演模型。表 6 给出了两个模型对测试集中每类云 CBH 的反演结果。

从表 6 可以看到：两个模型对 St/Sc、Cu 和 Ns 这 3 类云的反演效果相差较大，两个模型的 RMSE 和 MAE 之间的差距在 10^{1}～10^{2} m，R 之间的差距在 10^{-2}～10^{-1}，MRE 之间的差距在 10^{-2} 的量级；对其他 5 类云，两个模型的反演效果相差较小。对于 Ci、As 和 Ac，GBT 模型的 RMSE、MAE、R 和 MRE 均优于 RF 模型；对于 St/Sc、Ns 和 Dc，RF 模型的 RMSE、MAE、R 和 MRE 均优于 GBT 模型；对于 Cu 和 Multi 两类云，RF 模型的 RMSE、MAE 和 MRE 均优于 GBT 模型。据此，确定每类云最优的 CBH 反演模型，结果如表 7 所示。

4.2 方案二

与方案一确定模型参数的过程相同，先将决策树的最大深度设置为 3，再在训练验证集上训练模型，并根据样本的 RMSE 确定决策树数量的最佳取值。图 5 所示为训练验证集上样本的 RMSE 随两个模型决策树数量的变化。

表 6 各类云的两个模型在测试集上的反演效果

Table 6 Retrieval results of two models for all types of clouds on the test dataset

Cloud	Mean value /m	RF model				GBT model			
		RMSE /m	MAE /m	R	MRE /%	RMSE /m	MAE /m	R	MRE /%
Ci	9238.5	1060.6	841.4	0.7249	9.47	**1054.8**	**836.5**	**0.7285**	**9.41**
As	4918.6	1524.5	1180.9	0.6481	41.82	**1515.8**	**1173.7**	**0.6532**	**41.65**
Ac	2954.9	1357.2	1063.7	0.6396	66.68	**1352.2**	**1060.0**	**0.6431**	**66.51**
St/Sc	973.2	**620.0**	**372.5**	**0.6062**	**43.45**	669.5	394.4	0.5921	46.06
Cu	1075.6	**1000.3**	**634.4**	0.5018	**78.63**	1035.0	663.5	**0.5132**	84.22
Ns	1266.4	**1338.8**	**896.2**	**0.4246**	**106.35**	1371.7	933.2	0.4210	113.18
Dc	712.9	**534.5**	**369.5**	**0.1847**	**60.12**	538.4	375.6	0.1831	61.73
Multi	2285.0	**2044.7**	**1593.7**	0.4046	**148.92**	2044.9	1605.5	**0.4069**	150.72

表 7 每类云最优的 CBH 反演模型

Table 7 Optimal CBH retrieval model of all types of clouds

Cloud type	Retrieval model
Ci	GBT
As	GBT
Ac	GBT
St/Sc	RF
Cu	RF
Ns	RF
Dc	RF
Multi	RF

随着两个模型中决策树数量的增加，RMSE 的变化趋势基本一致，即 RMSE 先迅速减小后基本稳定。对于 RF 模型，将决策树的数量设为 13；对于 GBT 模型，将决策树的数量设为 163。根据设置的决策树数量重新在训练验证集上训练这两个模型，并根据样本的 RMSE 确定决策树最大深度的最佳取值。图 6 展示了训练验证集上样本的 RMSE 随两个模型中决策树最大深度的变化。

从图 6 可以看到，训练验证集上样本的 RMSE 随两个模型决策树最大深度的变化趋势类似。随着决策树最大深度的取值不断增加，RMSE 先迅速减小后逐渐增大。在整个变化过程中，RMSE 存在极小值。根据 RMSE 的极小值确定两个模型中决策树的最大深度。对于 RF 模型，将决策树的最大深度设为 11；对于 GBT 模型，将决策树的最大深度设为 5。根据确定的决策树的数量和最大深度，在训练验证集上对两个模型进行最终训练。利用测试集上的样本对两个模型的反演效果进行评估，结果如表 8 所示。

从表 8 的结果来看，两个模型在测试集上的反演效果相差较小，RMSE 和 MAE 的差距均在 5 m 以内，R 的差距在 0.001 以内，MRE 的差距在 0.3% 以内。RF 模型在测试集上的 MAE 和 MRE 两个指标均优于 GBT 模型，而 GBT 模型的 RMSE 和 R 两个指标则优于 RF 模型。分析这两个模型对测试集上各类云 CBH 的反演效果，结果如表 9 所示。

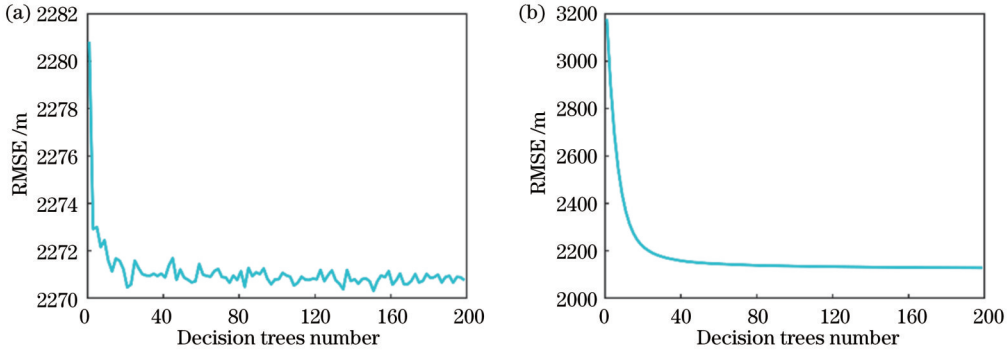

图 5 训练验证集上样本的 RMSE 随 RF 模型和 GBT 模型中决策树数量的变化。(a) RMSE 随 RF 模型中决策树数量的变化；(b) RMSE 随 GBT 模型中决策树数量的变化

Fig. 5 Variation of RMSE of samples on training and validation datasets with decision trees number in RF model and GBT model. (a) Variation of RMSE with decision trees number in RF model; (b) variation of RMSE with decision trees number in GBT model

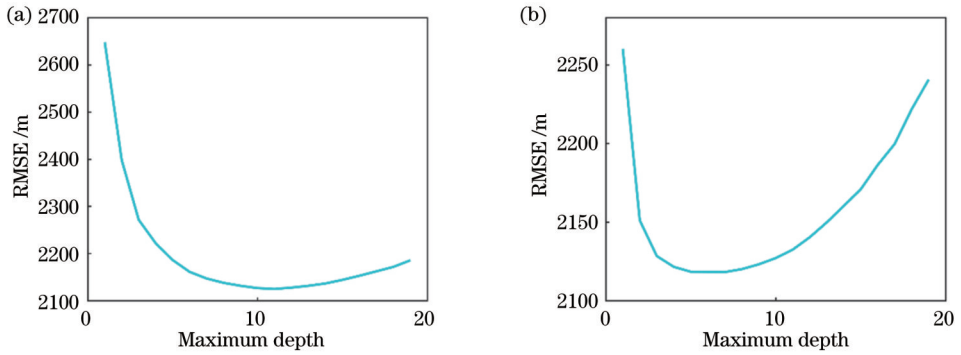

图 6 训练验证集上样本的 RMSE 随 RF 模型和 GBT 模型的决策树最大深度的变化。(a) RMSE 随 RF 模型的决策树最大深度的变化；(b) RMSE 随 GBT 模型的决策树最大深度的变化

Fig. 6 Variation of RMSE of samples on training and validation datasets with maximum depth of decision trees in RF model and GBT model. (a) Variation of RMSE with maximum depth of decision trees in RF model; (b) variation of RMSE with maximum depth of decision trees in GBT model

表 8 RF 模型和 GBT 模型在测试集上的反演效果

Table 8 Retrieval results of RF model and GBT model on test dataset

Model	RMSE /m	MAE /m	*R*	MRE /%
RF	2113.4	**1497.6**	0.7769	**124.60**
GBT	**2109.1**	1498.6	**0.7779**	124.81

从表 9 可以看到，两个模型对这 8 类云 CBH 的反演效果各有优劣，两种模型的整体效果也基本相当。综合来看，选择以 GBT 模型作为方案一最终的 CBH 反演模型。

4.3 方案比较

按照图 2 所示的反演方案，需要比较两种方案对 CBH 的反演效果，根据测试集上的结果选择最终的 FY-4A 云底高度反演模型。图 7 所示为两种方案的 CBH 反演模型对测试集上 129515 个样本的 CBH 反演结果。

从图 7 可以看到，两种方案的 CBH 模型反演结果均存在与参考线偏差较大的情况，例如在方案一的结果中，CloudSat 探测的 CBH 在 6～9 km，而模型反演的 CBH 却在 3 km 以下，这主要发生在多层云中。在现实中，多层云的组成非常复杂。CloudSat 总共有 8 种云类型，理论上共存在 56 种两层云的组合，而且现实中可能还有三层云甚至四层云的情况。相比之下，本文获取的多层云样本则不足以代表所有的情形。因此，基于这些样本训练的模型可能会得到与真实值相差较大的结果，模型还存在较大的改进空间。此外，CloudSat 探测的 CBH 在 3 km 以下的云，也存在模型反演结果与参考线偏差较大的情况，尤其是在方案二的反演结果中。方案二对这部分云 CBH 的反演结果在 5 km 以上的情况较多，Sc/St、Cu、Ns、Dc 以及多层云中均存在这种情况，这主要还是与方案二的反演思路有关，方案二没有区分云的类型，导致模型的反演结果与真实值相差较大。相比之下，方案一对这部分云反演的 CBH 大于 5 km 的情况明显少于方案二的反演结果。图 7 还列出了两种方案的 CBH 反演模型对测试集上样本的 RMSE、MAE、*R* 和 MRE。可以看到：方案一的 CBH 反演模型对测试集上所有样本的 RMSE 为 1304.7 m，MAE 为 898.3 m，*R* 为 0.9214，

表 9　RF 模型和 GBT 模型对测试集上各类云的 CBH 反演结果

Table 9　Retrieval results of RF model and GBT model for all types of clouds on test dataset

Cloud type	Mean value /m	RF model				GBT model			
		RMSE /m	MAE /m	R	MRE /%	RMSE /m	MAE /m	R	MRE /%
Ci	9238.5	**2528.7**	**2178.0**	0.6489	**23.80**	2529.7	2181.5	**0.6518**	23.84
As	4918.6	1896.3	1532.0	0.6111	41.19	**1891.8**	**1529.5**	**0.6140**	**41.04**
Ac	2954.9	1637.8	**1246.1**	**0.5566**	**63.07**	**1636.4**	1247.3	**0.5566**	63.17
St/Sc	973.2	1098.2	**536.6**	**0.2782**	**66.40**	**1095.6**	537.7	0.2763	66.69
Cu	1075.6	1734.9	1080.0	0.2450	171.41	**1725.9**	**1076.7**	**0.2458**	**170.77**
Ns	1266.4	1980.8	**1518.9**	**0.2353**	**246.69**	**1977.5**	1527.6	0.2320	248.25
Dc	712.9	3090.6	**2524.6**	−0.0628	**545.12**	**3071.9**	2532.3	**−0.0578**	546.44
Multi	2285.0	3104.5	2369.9	0.3775	**297.83**	**3095.7**	**2367.4**	**0.3783**	297.88

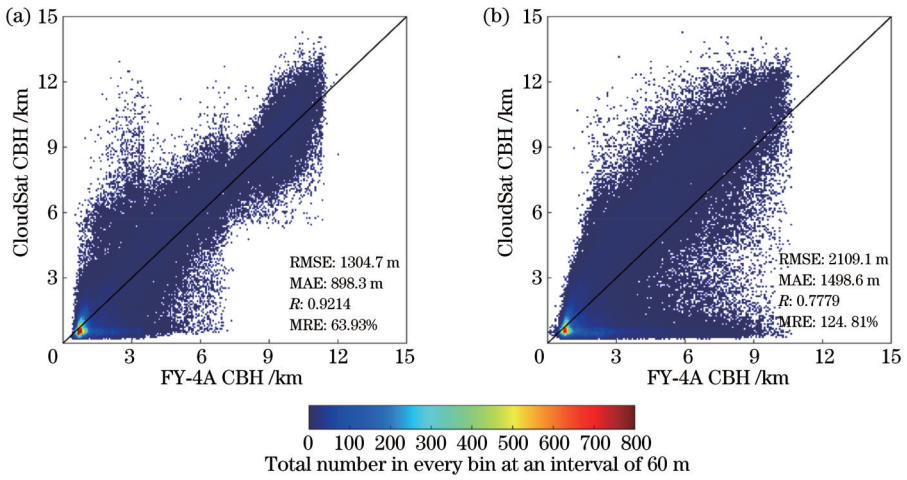

图 7　两种方案的模型对测试集样本的反演结果。(a)方案一;(b)方案二

Fig. 7　Retrieval results of models of two schemes on test dataset. (a) Scheme one; (b) scheme two

MRE 为 63.93%;方案二的 CBH 反演模型对测试集上所有样本的 RMSE 为 2109.1 m,MAE 为 1498.6 m,R 为 0.7779,MRE 为 124.81%。方案一的 4 个指标均优于方案二。此外,结合表 6 和表 9,得到了两种方案的 CBH 反演模型对每类云的 CBH 反演结果,如表 10 所示。

表 10　两种方案对测试集上每类云的 CBH 反演的结果

Table 10　Retrieval results of models of two schemes for all types of clouds on test dataset

Cloud	Mean value /m	Scheme one				Scheme two			
		RMSE /m	MAE /m	R	MRE /%	RMSE /m	MAE /m	R	MRE /%
Ci	9238.5	**1054.8**	**836.5**	**0.7285**	**9.41**	2529.7	2181.5	0.6518	23.84
As	4918.6	**1515.8**	**1173.7**	**0.6532**	41.65	1891.8	1529.5	0.6140	**41.04**
Ac	2954.9	**1352.2**	**1060.0**	**0.6431**	66.51	1636.4	1247.3	0.5566	**63.17**
St/Sc	973.2	**620.0**	**372.5**	**0.6062**	**43.45**	1095.6	537.7	0.2763	66.69
Cu	1075.6	**1000.3**	**634.4**	**0.5018**	**78.63**	1725.9	1076.7	0.2458	170.77
Ns	1266.4	**1338.8**	**896.2**	**0.4246**	**106.35**	1977.5	1527.6	0.2320	248.25
Dc	712.9	**534.5**	**369.5**	**0.1847**	**60.12**	3071.9	2532.3	−0.0578	546.44
Multi	2285.0	**2044.7**	**1593.7**	**0.4046**	**148.92**	3095.7	2367.4	0.3783	297.88

从表 10 可以看到:方案一的 CBH 反演模型对各类云 CBH 反演的 RMSE、MAE 和 R 均优于方案二的 CBH 反演模型,二者 RMSE 和 MAE 之间的差距普遍在 10^2～10^3 m,R 的差距在 10^{-1}～10^0;对于 MRE 这个指标,针对 As 和 Ac 外的其他 6 类云,方案一的 CBH 反演模型均优于方案二。这说明方案一 CBH 反演模型的效果明显优于方案二。因此,将方案一的 CBH 反演模型作为最终的 FY-4A 云底高度反演模型。

通过表 10 能够发现，两种方案的 CBH 反演模型对各类云的 CBH 反演效果存在较大的差异。对于 Ci 和 As 两类相对较高的云，模型反演结果与 CloudSat 参考值之间的 R 较大，MRE 较小，两种方案 CBH 反演模型的 R 均在 0.6 以上，MRE 均在 40% 以下，但方案二的 CBH 反演模型对 Ci 的 RMSE 和 MAE 较大，均大于 2000 m。对于 Dc，两种方案的 CBH 反演模型的 R 均最小。方案二的 CBH 反演模型的 R 甚至为负值，其 RMSE 和 MAE 均在 2500 m 以上，MRE 甚至超过了 500%；方案一的 CBH 反演模型对 Dc 的 R 只有 0.1847，但其 RMSE 和 MAE 均低于其他 7 类云，且 MRE 仅为 60%，远远低于方案二。另外，方案一的 CBH 反演模型对 Ac、St/Sc 和 Cu 的 R 在 0.5 以上，RMSE 和 MAE 基本等于或低于 1500 m，MRE 在 80% 以内。方案一的 CBH 反演模型对 Ns 和 Multilayer 的 R 也在 0.4 以上。

4.4　讨　　论

从 4.3 节可以看到，方案一对云底高度的反演效果明显更好。在反演云底高度时，首先对云的类型进行区分，然后对每类云分别采用独立的方法反演其云底高度，这种思路比不区分云的类型直接反演云底高度的思路更适合 FY-4A。

在实际应用中，可以按照图 8 所示的流程利用国家卫星气象中心提供的 FY-4A 的数据产品获取云底高度。由于 FY-4A 目前的云类型产品不能区分出上述 8 种类型的云，因此在运用所提出的云底高度反演方法时，需要根据 FY-4A 的数据产品区分出这 8 类云。可根据 Yu 等[32]提出的 FY-4A 云分类算法得到这 8 种云类型的分布，再对每类云分别反演其云底高度。具体来说：首先，根据 FY-4A 的反射率和亮温以及 CTH、COT 和 CER 等二级产品得到云类型；然后，结合 FY-4A 的 CTH、COT 和 CER 等产品，采用本文方法分别反演各类云的云底高度。因此，实际应用中云底高度可能的误差主要来自以下 3 个方面：1）云分类模型在对云进行分类时产生的误差。由于针对不同类型的云所采用的 CBH 反演模型不同，云分类的误差会导致后续 CBH 反演模型的选择出现偏差。2）FY-4A 的 CTH、COT 和 CER 等二级产品可能的偏差。虽然已经有学者证明了这些二级产品与 MODIS 和 Himawari-8 对应的产品有很好的相关性，但是随着 FY-4A 在轨运行时间的延长，仪器本身可能会出现一些问题，或者受到噪声的干扰，导致相关的产品产生误差[33]。3）CBH 反演模型本身对云底高度的反演误差。从 4.3 节对反演模型的评估来看，与 CloudSat 探测的云底高度相比，本文提出的 CBH 反演模型还有改进的空间。

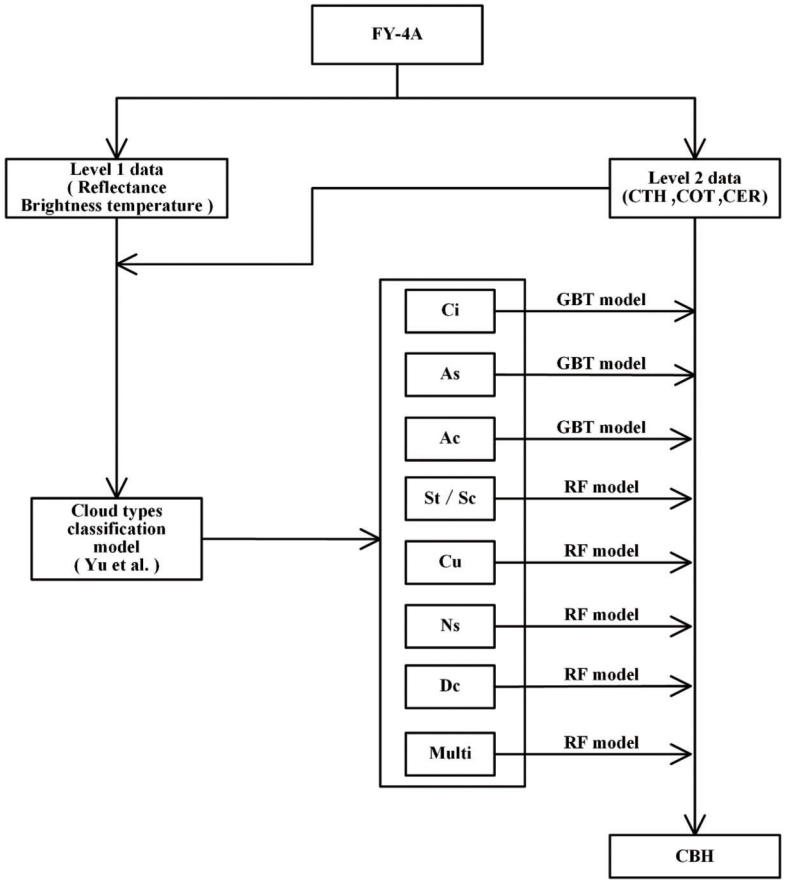

图 8　实际应用中的 FY-4A 云底高度反演流程

Fig. 8　Retrieval flow of CBH for FY-4A in practical application

4.5　个例分析

图 9 给出了一个具体的例子来展示如何利用 FY-4A/AGRI 的数据产品获取云底高度。在该例子中，FY-4A 的时间是 2017 年 11 月 20 日 5 时 45 分至 5 时 59 分（世界时，后同），选取临近时刻 CloudSat 的云底高度作为对比。CloudSat 此次扫描的起始时间是 2017 年 11 月 20 日 5 时 1 分，CloudSat 经过图 9 区域的时间大致为 5 时 53 分至 6 时 1 分（实线为 CloudSat 的轨迹）。

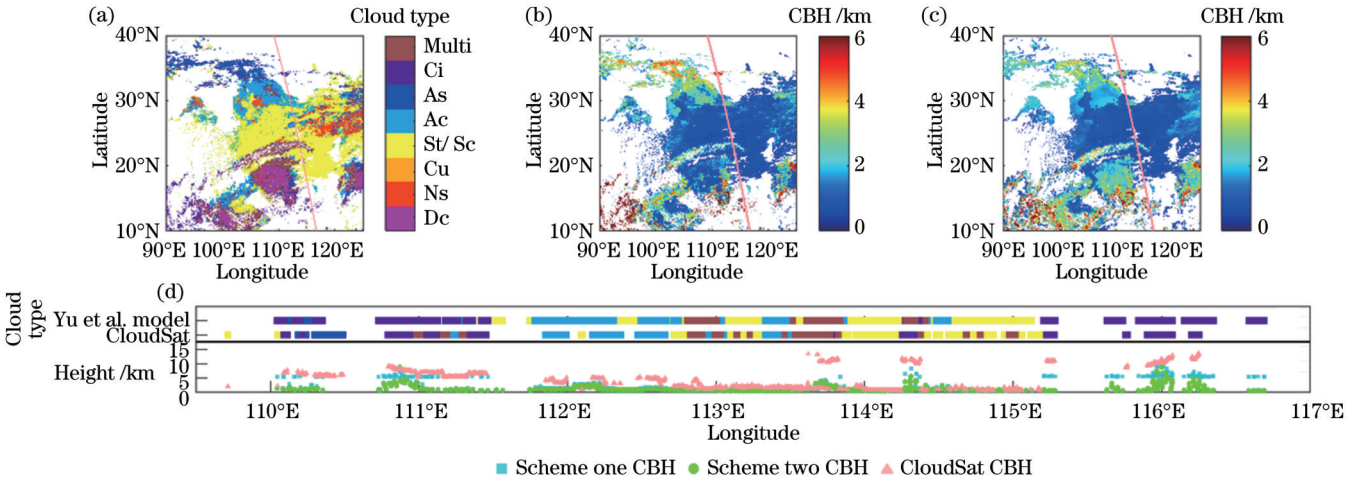

图 9　两种方案的模型反演的云底高度及其与 CloudSat 探测结果的比较。(a)根据文献[32]提出的云分类模型得到的云类型；(b)根据图 9(a)的云分类结果和方案一的模型反演的云底高度；(c)方案二的模型反演的云底高度；(d)CloudSat 轨迹上，CloudSat 探测的云类型与根据文献[32]的云分类模型得到的云类型，以及两种方案的模型反演的云底高度与 CloudSat 探测的云底高度

Fig. 9　CBH retrieved from models of two schemes and the comparison with CBH from CloudSat. (a) Cloud types obtained according by the model proposed in Ref. [32]; (b) CBH retrieved from the cloud types of Fig. 9(a) and the model of scheme one; (c) CBH retrieved from the model of scheme two; (d) comparison between the cloud types of CloudSat and the model proposed in Ref. [32], and comparison among CBH retrieved from models of two schemes and CBH from CloudSat on CloudSat track

根据 Yu 等[32]提出的 FY-4A 云分类模型得到这 8 种云类型，如图 9(a)所示；根据图 9(a)的云类型和方案一的 CBH 反演模型得到云底高度，如图 9(b)所示；图 9(c)所示为根据方案二的 CBH 反演模型得到的云底高度；图 9(d)展示了在 CloudSat 轨迹上，根据 Yu 等[32]提出的 FY-4A 云分类模型得到的云分类结果和 CloudSat 探测的云类型，以及两种方案反演的云底高度和 CloudSat 探测的云底高度。

从图 9(b)、(c)可以看到：两种方案的模型在 20°N～30°N 范围内对云底高度的反演结果比较接近，这一范围内的云底高度在 1 km 左右；在 32°N 以北范围内，方案一的模型反演的云底高度比方案二更高，在 3 km 以上；在 20°N 以南，方案二的模型反演的云底高度比方案一更高；在 17°N、110°E 周围以及 16°N～20°N、122°E～124°E 范围内，方案二的模型反演的云底高度比方案一的模型反演的云底高度高出约 1 km。如图 9(d)所示，在 CloudSat 轨迹上，两种方案的模型反演的云底高度与 CloudSat 探测的云底高度之间均存在一定的差异。对于云底高度在 10 km 以上的云，两种方案的模型反演的云底高度普遍低于 CloudSat 探测的云底高度，尤其是在 116°E 附近，CloudSat 探测到的云底高度普遍在 10 km 以上，方案二的模型反演的云底高度明显低于方案一的模型反演的结果，而方案一的模型反演的云底高度与 CloudSat 探测的云底高度之间的差值普遍在 3 km 以上，相差较大。对 6 km 及以下的云，两种方案的模型反演的云底高度与 CloudSat 探测的云底高度相差较小。例如在 113°E～115°E 范围内，方案一的模型反演的云底高度与 CloudSat 探测的云底高度相差较小，方案二的模型反演的云底高度略低于方案一的模型反演的云底高度。

从图 9 可以看到，对 CBH 在 10 km 以上的云（在上述例子中，CloudSat 对这些云的分类为 Ci），模型反演结果与 CloudSat 探测结果相差较大。就云类型而言，除了 114.3°E 附近 CBH 在 10 km 以上的云，对其他 CBH 在 10 km 以上的云，根据 Yu 等[32]提出的云分类模型得到的云类型结果与 CloudSat 探测结果基本一致。因此，在该例子中，对 CBH 在 10 km 以上的云，模型反演的云底高度与 CloudSat 探测的云底高度相差较大的原因主要是 CBH 反演模型本身，在该例子中，两种方案的模型对 Ci 的 CBH 反演误差较大。这个例子说明两种方案 CBH 反演模型结果的精度还可以进一步提高，可以对模型做进一步的优化。在现有模型中，输入数据只是 FY-4A/AGRI 的 3 种云产品，输入的数据较少。在后续研究中，可以考虑将再分析资料和 FY-4A/AGRI 光谱通道的反射率和亮度、温度作为输入数据增加到模型中，进一步训练模型，对模型进行优化。

5 结 论

云底高度作为云的重要边界参数，不但影响地气系统的辐射收支，还影响着飞行活动的安全。由于目前静止气象卫星没有云底高度的业务产品，本文基于集成学习的理论提出针对FY-4A卫星的云底高度反演方法，设计了云底高度的两种反演方案，希望找到更加适合FY-4A的云底高度反演思路，为后续静止气象卫星云底高度业务产品的开发提供参考。

第一种方案先区分云的类型，再对每类云分别采用独立的集成学习模型反演其云底高度；第二种方案不区分云的类型，采用统一的集成学习模型反演云底高度。以CloudSat探测的云底高度对两种方案的反演结果进行对比分析：方案一反演模型的RMSE为1304.7 m，MAE为898.4 m，R为0.9214，MRE为63.93%；方案二反演模型的RMSE为2109.1 m，MAE为1498.6 m，R为0.7779，MRE为124.81%。方案一的反演效果明显优于方案二。因此在反演云底高度时，先对云的类型进行区分，再对每类云分别采用独立的方法反演其云底高度，这种思路比不区分云的类型直接反演云底高度的思路效果更好。

此外，所提出的云底高度反演模型仍然还有改进的空间，对于某些类型的云，模型反演的效果还有待提高。例如，模型对深对流云云底高度反演的RMSE和MAE分别为534.5 m和369.5 m，虽然这两个指标在8类云中最小，但是其R仅为0.1847，同样也是8类云中的最小值。在后续研究中，可以考虑将再分析资料或者FY-4A/AGRI的一级数据加入到云底高度反演模型中，以提升模型的准确性。

参 考 文 献

[1] Viúdez-Mora A, Costa-Surós M, Calbó J, et al. Modeling atmospheric longwave radiation at the surface during overcast skies: the role of cloud base height[J]. Journal of Geophysical Research: Atmospheres, 2015, 120(1): 199-214.

[2] Herzegh P, Wiener G, Bateman R, et al. Data fusion enables better recognition of ceiling and visibility hazards in aviation[J]. Bulletin of the American Meteorological Society, 2015, 96(4): 526-532.

[3] 严卫, 韩丁, 周小珂, 等. 利用CloudSat卫星资料分析热带气旋的结构特征[J]. 地球物理学报, 2013, 56(6): 1809-1824.
Yan W, Han D, Zhou X K, et al. Analysing the structure characteristics of tropical cyclones based on CloudSat satellite data[J]. Chinese Journal of Geophysics, 2013, 56(6): 1809-1824.

[4] Kahn B H, Chahine M T, Stephens G L, et al. Cloud type comparisons of AIRS, CloudSat, and CALIPSO cloud height and amount[J]. Atmospheric Chemistry and Physics, 2008, 8(5): 1231-1248.

[5] 尚华哲, 胡斯勒图, 李明, 等. 基于被动遥感卫星可见至红外通道观测的云特性遥感[J]. 光学学报, 2022, 42(6): 0600003.
Shang H Z, Husi L T, Li M, et al. Remote sensing of cloud properties based on visible-to-infrared channel observation from passive remote sensing satellites[J]. Acta Optica Sinica, 2022, 42(6): 0600003.

[6] Hutchison K D. The retrieval of cloud base heights from MODIS and three-dimensional cloud fields from NASA's EOS Aqua mission[J]. International Journal of Remote Sensing, 2002, 23(24): 5249-5265.

[7] Liou K N. Radiation and cloud processes in the atmosphere: theory, observation and modeling[M]. New York: Oxford University Press, 1992.

[8] Seaman C J, Noh Y J, Miller S D, et al. Cloud-base height estimation from VIIRS. part I: operational algorithm validation against CloudSat[J]. Journal of Atmospheric and Oceanic Technology, 2017, 34(3): 567-583.

[9] Noh Y J, Forsythe J M, Miller S D, et al. Cloud-base height estimation from VIIRS. part Ⅱ: a statistical algorithm based on a-train satellite data[J]. Journal of Atmospheric and Oceanic Technology, 2017, 34(3): 585-598.

[10] Forsythe J M, Haar T H V, Reinke D L. Cloud-base height estimates using a combination of meteorological satellite imagery and surface reports[J]. Journal of Applied Meteorology, 2000, 39(12): 2336-2347.

[11] 王帅辉, 姚志刚, 韩志刚, 等. CloudSat云底高度外推估计的可行性分析[J]. 气象, 2012, 38(2): 210-219.
Wang S H, Yao Z G, Han Z G, et al. Feasibility analysis of extending the spatial coverage of cloud-base height from CloudSat[J]. Meteorological Monthly, 2012, 38(2): 210-219.

[12] 李浩然, 孙学金, 刘磊, 等. 基于模板匹配的云底高度估计[J]. 气象科学, 2015, 35(5): 610-615.
Li H R, Sun X J, Liu L, et al. Cloud base height estimation based on template matching[J]. Journal of the Meteorological Sciences, 2015, 35(5): 610-615.

[13] 高顶. 基于FY-4A卫星的云底高度反演与应用研究[D]. 长沙: 国防科技大学, 2018.
Gao D. Research on cloud base height retrieval and application based on FY-4A[D]. Changsha: National University of Defense Technology, 2018.

[14] 谭仲辉, 马烁, 韩丁, 等. 基于随机森林算法的FY-4A云底高度估计方法[J]. 红外与毫米波学报, 2019, 38(3): 381-388.
Tan Z H, Ma S, Han D, et al. Estimation of cloud base height for FY-4A satellite based on random forest algorithm[J]. Journal of Infrared and Millimeter Waves, 2019, 38(3): 381-388.

[15] Tan Z H, Huo J, Ma S, et al. Estimating cloud base height from Himawari-8 based on a random forest algorithm[J]. International Journal of Remote Sensing, 2021, 42(7): 2485-2501.

[16] Lin H, Li Z L, Li J, et al. Estimate of daytime single-layer cloud base height from advanced baseline imager measurements [J]. Remote Sensing of Environment, 2022, 274: 112970.

[17] 黄鹏宇, 郭强, 韩昌佩, 等. FY-4A/GIIRS资料云上温度廓线反演研究[J]. 激光与光电子学进展, 2021, 58(17): 1701002.
Huang P Y, Guo Q, Han C P, et al. Research on retrieval of temperature profile on cloud based on FY-4A/GIIRS data[J]. Laser & Optoelectronics Progress, 2021, 58(17): 1701002.

[18] Xu W J, Lü D R. Evaluation of cloud mask and cloud top height from Fengyun-4A with MODIS cloud retrievals over the Tibetan Plateau[J]. Remote Sensing, 2021, 13(8): 1418.

[19] Lai R Z, Teng S W, Yi B Q, et al. Comparison of cloud properties from Himawari-8 and FengYun-4A geostationary satellite radiometers with MODIS cloud retrievals[J]. Remote Sensing, 2019, 11(14): 1703.

[20] 崔林丽, 郭巍, 葛伟强, 等. FY-4A卫星云顶参数精度检验及台风应用研究[J]. 高原气象, 2020, 39(1): 196-203.
Cui L L, Guo W, Ge W Q, et al. Comparisons of cloud top parameter of FY-4A satellite and its typhoon application research [J]. Plateau Meteorology, 2020, 39(1): 196-203.

[21] 余茁夫, 马烁, 胡雄, 等. 基于多源数据的"利奇马"台风大气环流、云及降水特征分析[J]. 气象科学, 2020, 40(1): 41-52.
Yu Z F, Ma S, Hu X, et al. Analysis of atmospheric circulation, cloud and precipitation characteristics of typhoon "Lekima" based on multi-source data[J]. Journal of the

Meteorological Sciences, 2020, 40(1): 41-52.
[22] 王清平, 朱雯娜, 王勇, 等. FY-4A资料在乌鲁木齐机场浓雾天气监测中的初步应用[J]. 气象, 2021, 47(5): 627-637.
Wang Q P, Zhu W N, Wang Y, et al. Preliminary application of FY-4A satellite data in dense fog weather events at Urumqi international airport[J]. Meteorological Monthly, 2021, 47(5): 627-637.
[23] 袁锦涵, 周永波, 刘玉宝, 等. 云滴谱分布对FY-4A/AGRI水云光学厚度与有效粒子半径反演的影响[J]. 光学学报, 2022, 42(6): 0628004.
Yuan J H, Zhou Y B, Liu Y B, et al. Effect of cloud droplet spectrum distribution on retrievals of water cloud optical thickness and effective particle radius by AGRI onboard FY-4A satellite[J]. Acta Optica Sinica, 2022, 42(6): 0628004.
[24] Lao P, Liu Q, Ding Y H, et al. Rainrate estimation from FY-4A cloud top temperature for mesoscale convective systems by using machine learning algorithm[J]. Remote Sensing, 2021, 13(16): 3273.
[25] Liu P, Yang Y, Gao J D, et al. An approach for assimilating FY4 lightning and cloud top height data using 3DVAR[J]. Frontiers in Earth Science, 2020, 8: 288.
[26] Wang Z E. 2019. Level 2 cloud scenario classification product process description and interface control document [EB/OL]. (2019-05-17)[2022-04-10]. https://www.cloudsat.cira.colostate.edu/cloudsat-static/info/dl/2b-cldclass/2B-CLDCLASS_PDICD.P1_R05.rev1_.pdf.
[27] Min M, Li J, Wang F, et al. Retrieval of cloud top properties from advanced geostationary satellite imager measurements based on machine learning algorithms[J]. Remote Sensing of Environment, 2020, 239: 111616.
[28] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.
[29] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
[30] 李铸, 张庆永, 孔令华, 等. 基于激光诱导击穿光谱与随机森林识别GCr15钢的硬度[J]. 中国激光, 2022(9): 0911002.
Li Z, Zhang Q Y, Kong L H, et al. Hardness characterization of GCrl5 steel based on laser-induced breakdown spectroscopy and random forest[J]. Chinese Journal of Lasers, 2022(9): 0911002.
[31] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.
[32] Yu Z F, Ma S, Han D, et al. A cloud classification method based on random forest for FY-4A[J]. International Journal of Remote Sensing, 2021, 42(9): 3353-3379.
[33] 李文力, 李秀举, 屠黄唯, 等. 像元间光谱响应非均匀性与条带噪声的关系[J]. 光学学报, 2022, 42(12): 1211001.
Li W L, Li X J, Tu H W, et al. Relationship between spectral response non-uniformity of pixels and stripe noise[J]. Acta Optica Sinica, 2022, 42(12): 1211001.

Cloud Base Height Retrieval Methods for FY-4A Based on Ensemble Learning

Yu Zhuofu[1], Wang Ya[2*], Ma Shuo[1**], Ai Weihua[1], Yan Wei[1]

[1]*College of Meteorology and Oceanography, National University of Defense Technology, Changsha* 410000, *Hunan, China*;

[2]*National Satellite Meteorological Center, China Meteorological Administration, Beijing* 100081, *China*

Abstract

Objective Cloud base height (CBH) is a crucial cloud parameter affecting the water cycle and radiation budget of the earth-atmosphere system. Additionally, CBH has a great impact on aviation safety. Low CBH often leads to a decrease in visibility, which poses a great threat to flight safety. Therefore, it is meaningful to acquire accurate CBH for related scientific research and meteorological services. It is valuable but challenging to use satellite passive remote sensing data to retrieve CBH. Some cloud products such as cloud top height (CTH) and cloud optical thickness (COT) are often used in previous research, related to CBH retrieval, from which two ideas to retrieve CBH can be summarized. The first idea employed independent methods to obtain CBH of different types of clouds respectively, and the second one directly retrieves CBH using cloud products of satellites without regarding cloud types. At present, there is no CBH products of FY-4A. Therefore, a CBH retrieval method for FY-4A is introduced in this paper. According to the two ideas mentioned above, two schemes of CBH retrieval are designed, which are compared to find more suitable ideas to retrieve CBH for FY-4A and to provide reference for subsequent development of FY-4A CBH products.

Methods A CBH retrieval method based on ensemble learning is proposed in this paper. CTH, COT, and cloud effective radius (CER) from FY-4A are used. Additionally, CBH and cloud types from CloudSat are employed for their widely recognized data quality. First, data of FY-4A and CloudSat are matched spatiotemporally and are divided into training data, validation data, and test data. Second, CBH retrieval models are built based on two ensemble learning algorithms, random forest (RF), and gradient boosting tree (GBT). Two schemes of CBH retrieval are designed in this paper. In the first scheme, matched data are divided into eight types according to the eight cloud types of CloudSat. For each type of cloud, two retrieval models are built based on RF and GBT using training data and validation data through ten-

fold cross validation. The optimal model is selected according to the models′ results on test data. In the second scheme, retrieval models are built without regarding cloud types. Training data of the eight cloud types are combined together. Validation data and test data are processed similarly. The three data sets are used to obtain the RF model and GBT model, and to select the optimal retrieval model. Finally, the optimal scheme and model of CBH retrieval for FY-4A are selected according to the models′ performance.

Results and Discussions Root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient (R), and mean relative error (MRE) are used to evaluate models′ performance. In the first scheme, the GBT model is the optimal retrieval model for Cirrus (Ci), Altostratus (As), and Altostratus (Ac). RF model is the optimal retrieval model for Stratus/Stratocumulus (St/Sc), Cumulus (Cu), Nimbostratus (Ns), deep convective cloud (Dc), and multilayer cloud (Multi). In the second scheme, the GBT model is the optimal retrieval model. The models of the two schemes are compared on test data with 129515 samples. Overall, the retrieval model of the first scheme outperforms that of the second scheme. Specifically, RMSE of the model in the first scheme is 1304.7 m. MAE is 898.3 m, R is 0.9214, and MRE is 63.93%. For the eight types of clouds, RMSE, MAE, R, and MRE of the model in the first scheme are also superior to those of the model in the second scheme. Although the first scheme can obtain better results, the retrieval model of the first scheme still needs to be improved in the future. For example, the performance of the retrieval model for Dc is not a patch on that of other types of clouds. Additionally, the paper discusses how to apply the proposed method to practice. First, level 1 data (i.e. reflectance and brightness temperature) and level 2 data (i.e. CTH, COT, and CER) of FY-4A can be used to acquire the eight cloud types according to a cloud type classification model proposed by Yu et al. Second, according to the cloud type classification results, the retrieval models of the first scheme can be adopted to retrieve CBH for the eight types of clouds respectively.

Conclusions CBH is a critical cloud parameter, but there are no CBH products of geostationary meteorological satellites currently. Thus, a CBH retrieval method for FY-4A based on ensemble learning is introduced in this paper. Two schemes of CBH retrieval are designed, and corresponding CBH retrieval models are built based on two ensemble learning algorithms, namely, RF and GBT. Data of CTH, COT, and CER from FY-4A are used in this paper. The first scheme employs eight independent models to retrieve CBH for eight types of clouds (i.e. Ci, As, Ac, St/Sc, Cu, Ns, Dc, and Multi) respectively. Specifically, for Ci, As, and Ac, the GBT model is used to retrieve CBH. For the other five types of cloud, the RF model is used to retrieve CBH. The second scheme uses a GBT model to retrieve CBH without regarding cloud types. CBH from CloudSat is used to evaluate the results of the two schemes, and the retrieval model of the first scheme outperforms that of the second scheme. For the eight types of clouds, the retrieval model of the first scheme also obtains better results.

Key words atmospheric optics; cloud base height retrieval; FY-4A; cloud top height; cloud optical thickness; cloud effective radius; ensemble learning