基于多角度的日间北极云检测卫星数据与案例研究

石涛, 余斌, Eugene E. CLOTHIAUX, and Amy J. BRAVERMAN

全球气候模型预测，地表空气温度对大气中二氧化碳水平增加的依赖性最强，将出现在北极地区。对这些依赖性的系统研究需要准确的北极范围的测量，特别是云的覆盖。因此，北极地区的云层探测是非常重要的，但由于云层和冰雪覆盖的表面具有相似的遥感特征，因此也是一个挑战。本文利用多角度成像光谱辐射计（MISR）图像提出了两种新的实用的北极云层探测算法。其关键思想是识别图像中的无云表面像素，而不是像现有的MISR操作算法中的有云像素。通过广泛的探索性数据分析和使用领域知识，已经确定了三个物理上有用的特征来区分表面像素和多云像素。第一种算法，增强型线性相关匹配（ELCM），用固定的或数据自适应的截止值对这些特征进行阈值处理。通过使用ELCM标签作为Fisher二次判别分析（QDA）的训练数据来获得概率标签，从而形成第二种（ELCM-
QDA）算法。这两种算法都是自动化的，而且计算效率高，可以对大规模的MISR数据集进行操作处理。基于500万个专家标记的像素，ELCM的结果在准确率（92%）和覆盖率（100%）方面与两个MISR操作算法相比都很明显，一个准确率为80%，覆盖率为27%；另一个准确率为83%，覆盖率为70%。ELCM-
QDA的概率预测也与专家的标签一致，信息量更大。总之，ELCM和ELCM-
QDA在使用MISR数据的所有可用操作算法中提供了迄今为止最好的性能。

关键字：分类；聚类；特征选择；多角度成像光谱仪；二次判别分析；再遥感。

1. 简介

地球气候对大气层二氧化碳量增加的敏感性是一个普遍和科学界感兴趣的话题，也是一个重要的公共政策问题。今天的全球气候模型普遍预测，在整个21世纪，如果大气中的二氧化碳水平增加一倍，全球表面空气温度将增加1.5-
3.5K。这些模型还预测，地表空气温度对大气二氧化碳水平增加的依赖性最强，将发生在北极地区（Giorgi和Bi 2005）。随着北极的变暖，冰雪覆盖的表面、大气中的水蒸气和云层的属性和分布的变化可能会导致更多的变暖，从而对大气中的二氧化碳量的增加产生强烈的敏感性。对这些问题的系统研究需要在北极范围内进行准确的测量（例如，Francis, Hunter, Key和Wang，2005），特别是云层的测量，因为云层在调节二氧化碳浓度方面起着重要作用。

石涛是俄亥俄州立大学统计系助理教授，俄亥俄州哥伦布市，43210（电子邮件：*taoshi@stat.osu.edu*）。Bin Yu是加州大学统计系教授，Berkeley, CA 94720（电子邮件：*binyu@stat.berkeley.edu*）。Eugene E. Clothiaux是宾夕法尼亚州立大学气象学系副教授，宾夕法尼亚州立大学公园，16802（电子邮件：*cloth@essc.psu.edu*）。Amy J. Braverman是加州理工学院喷气推进实验室信息系统和计算机科学部的统计员和高级成员，帕萨迪纳，加州91109（电子邮件：*Amy.Braverman@jpl.nasa.gov*）。这项研究得到了美国国家科学基金会FD01-12731（Yu）、CCR-0106656（Shi和Yu）、DMS-03036508（Shi和Yu）、DMS-0605165（Yu）以及ARO资助W911NF-05-1-
0104（Yu）的部分支持。这也是2004年春季Yu获得的Miller基金会教授职位的部分支持。Clothiaux得到了美国国家航空和航天局（NASA）资助NNG04GL93G和喷气推进实验室合同1259588的支持。Braverman的工作是在加州理工学院喷气推进实验室进行的，与NASA签有合同。MISR数据是由NASA兰利研究中心大气科学数据中心提供的。第4.2节中的结果是由加州大学伯克利分校于斌的博士生Guilherme Rocha提供的。作者感谢L. Di Giro-lamo、D. J. Diner、R. Davies和R. Kahn的有益讨论和建议，特别是感谢喷气推进实验室的Dominic Mazzoni分享、调整和专业支持他的软件包MISRLEARN供我们在这项研究中使用。本研究中所有的专家标签数据都是亲手制作的。

北极对地表空气温度上升的敏感性（例如，Kato等人，2006）。

确定北极地区云的特性是一个棘手的问题，因为液态和冰态的云颗粒往往具有与构成冰雪覆盖表面的颗粒类似的散射特性。因此，从云层和冰雪覆盖的表面发出的可见和红外电磁辐射量往往是相似的，这导致了在这些表面类型上探测云层的问题。如果不准确描述北极上空的云层，我们就无法评估它们对通过北极大气层的太阳能和地面电磁辐射流的影响，也无法确定它们的变化是否会加强或改善未来北极的变暖。

随着1999年美国国家航空航天局(NASA)Terra卫星上的多角度成像光谱辐射仪(MISR)的发射，在9个视角下进行的新型电磁辐射测量成为了一种新的方法。

可用于科学研究。不同于传统的多光谱

使用他的MISRLEARN软件包得出的数据。这篇文章的完整版本可在网上查阅：*www.stat.berkeley.edu/~binyu/publications.html*。

与在单一视角下进行测量的传感器不同，MISR传感器由九台摄像机组成（图1），每台摄像机在四个光谱带（蓝、绿、红和近红外）的不同角度观察地球场景。这九台摄像机的天顶角为
摄像机前向为70.5° （Df），60.0° （Cf），45.6° （Bf），26.1° （Af）；天底方向为0.0° （An），尾向为26.1° （Aa），45.6° （Ba），60.0° （Ca），70.5° （Da）。(摄像机的字母名称中的 "f "是指代表 "前进 "方向，"a "代表 "船尾 "方向）。Da相机在Df相机的7分钟后收集数据。

MISR产生了大量的数据，因为它以高空间分辨率覆盖了全球。MISR相机覆盖了地球表面约360公里宽的范围，从北极到地球的日光面。

计学会杂志
2008年6月，第103卷，第482期，应用和案例研究
doi 10.1198/016214507000001283

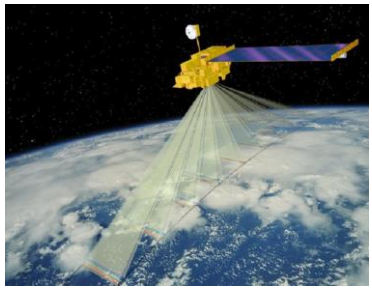

图1.Terra卫星的卡通图，其中有九个MISR相机的视角方向。图片由喷气推进实验室的MISR科学团队提供。

到南极洲大约需要45分钟。有233个地理上不同但重叠的MISR区域，这些区域也被称为*路径*。MISR以16天的重复周期收集所有路径的数据；也就是说，它每16天覆盖一次完全相同的路径。在MISR的数据产品中，每条路径被细分为180个*区块*，区块编号从北极到南极增加。MISR的每一次完整行程都围绕着
地球被赋予了它自己的*轨道*号。每个MISR像素覆盖了地面上275米×275米的区域，由于MISR的全球覆盖，产生了大量的数据。例如，一个轨道的数据是巨大的--
MISR平均每秒收集3.3兆比特，峰值速率为9.0兆比特/秒
。
由于传输通道的限制，这些大量的数据集不能被下行链路；因此，只有来自天底相机的红色辐射和所有通道以275米的分辨率传输，而来自非天底相机的蓝色、绿色和近红外辐射则在机上汇总为一个完整的数据。
1.1 公里×1.1公里的分辨率，然后再传输。
MISR的角度辐射值代表了一种全新的
从一开始，他们就表现出云层与冰雪覆盖的表面可能分离的强烈迹象（Diner等人，1999a）。然而，MISR的云层探测算法（见第2.1节中的细节）是在MISR启动之前设计的，它们并不特别针对极地地区明亮表面的云层探测。因此，现有的MISR运行算法在极地地区不能很好地工作。此外，庞大的数据集规模对任何算法都构成了障碍，并对极地云层探测问题的候选统计方案施加了严重的计算限制。在这种情况下，MISR数据的巨大规模是决定统计方法是否可行的一个非常重要的因素，所选择的方法必须包括非统计学、科学和操作方面的考虑。

北极地区的云层探测不仅在科学上很重要，而且在统计学上也很有挑战性。标准的分类框架并不容易适用，因为考虑到数据的大小和资源时间的限制，不可能在操作上获得专家标签。即使在离线情况下，最先进的分类器，如高斯核支持向量机（Vapnik，1995），在离线专家标签上进行训练，在应用于MISR辐射度测量时也不能提供令人满意的精度（参见第4.1节）。聚类框架也不合适，因为每个数据单元（本文中定义为每组连续的三个MISR块）都可能是
完全被云覆盖或完全无云。当一个数据单元完全被云覆盖时，聚类对应于不同类型的多云像素，但不对应于一个像素是否多云或清晰。因此，统计学上的挑战是如何以一种高效的计算方式将分类和聚类方案结合起来。

这项工作的目标是建立可操作的云检测算法，能够在不需要人类干预（专家标注）的情况下，一次有效地处理大规模的MISR数据集。我们利用广泛的探索性数据分析的结果以及与MISR科学团队的互动来构建一个标签方案，或称检测器。对于每个数据单元，我们将当前数据单元的聚类信息与从以前数据单元学到的检测器结合起来。因此，该算法是连续的，而且它自适应地结合了分类和聚类方法。

我们方法上的创新之处在于寻找*无云*的而不是冰雪覆盖的地表图像像素。目前实用的MISR算法使用多种方法寻找有云的像素，包括两个不同的MISR相机之间的平行度和亮度差异。相比之下，我们的方法是利用同一场景在无遮挡、无云条件下固有的多个视图之间的亮度相关性。我们表明，这种检测目标的简单逆转确实是富有成效的。这也是一个典型的例子，说明统计学的洞察力可以为传统上以非统计学方式解决的问题带来新的工具。

在我们的聚类和分类框架中，对地表的建模比对云层的建模更有优势，因为地表在不同的视角之间不会有太大的变化，而云层在不同的视角之间总是不同的。我们的算法是基于三个物理上有用的特征：用于描述冰雪覆盖表面的散射特性的不同MISR观测方向的同一场景的MISR图像的相关性（CORR），整个场景的MISR天底相机像素值的标准偏差（SD_{An}
），以及归一化差异角度指数（NDAI），它描述了场景中MISR观测方向的变化。经过统计学家和大气科学家的大量探索性数据分析，我们得出了这组特征。

我们提出了一种增强的线性相关匹配（ELCM）算法，该算法基于对三个特征的固定值或数据适应性的阈值。我们设想ELCM算法将依次应用于MISR数据单元。CORR和SD_{An}
特征的截止值，或阈值，在操作处理过程中被设置为固定值，因为它们在不同时间和地点收集的数据中是稳定的。NDAI阈值最初是根据对第一组数据单元的离线分析设定的。为了对随后的数据单元进行分类，NDAI阈值要么保持在以前的值，要么使用适用于当前数据单元的数据适应性算法进行更新。ELCM算法产生的标签然后被用来训练Fisher的二次判别分析（QDA）以产生概率标签，从而形成我们的第二个（ELCM-
QDA）检测器。QDA步骤产生了 "无云"的概率，作为一个更有信息量的概率预测。

为了评估ELCM和ELCM-
QDA算法的准确性，它们被应用于广泛的MISR数据集和

与专家标签进行测试。专家标签是为2002年白天在北极上空收集的10个轨道的MISR数据提供的，它们代表了足够数量的最佳可用估值数据（即508.6万个1.1公里的像素）。

文章的其余部分组织如下。第2节包含对MISR算法的简要回顾，并介绍了本研究中调查的MISR数据。第3节首先描述了我们提出的方法，从研究的理由和特征的计算开始，然后提出了我们基于特征自适应阈值的ELCM算法（第3.2节），并证明了通过使用ELCM标签来训练Fisher的QDA，可以获得对部分阴天场景的概率预测（第3.3节）。第4节介绍了测试结果。ELCM和ELCM-QDA的结果与专家标签以及MISR操作算法进行了比较，探讨了我们算法的优势和劣势，最后，专家标签被用来说明我们的特征比原始辐射测量值更适合于分离晴天和阴天的像素。第5节以总结和讨论我们算法的潜在科学影响来结束文章。

2. Misr操作算法和数据描述

从多个角度观察大气层和地表，MISR具有立体能力，可用于检索地球表面或上方的物体，如云层的延伸。除了立体信息外，MISR的非天底相机角度提供了不同物体（如云）的辐射散射模式。这两个新的信息来源促使MISR的2级大气层顶部云（L2TC）算法（Diner等人，1999b）产生了两个云掩码：立体衍生云掩码（SDCM）和角标云掩码（ASCM）。在第2.1节中，我们简要回顾了MISR的运行算法和它们在极地地区的缺陷。在第2.2节中，我们描述了本研究中使用的数据。

2.1 MISR操作算法

L2TC算法利用MISR的立体能力，通过比较检索到的物体高度和下面的已知地形高度来检测云层。L2TC云层高度检索算法的原理是基于MISR测量值与已知的基于地球的参考椭球体的登记。如图2所示，一朵云被登记到不同的参考椭圆体上。

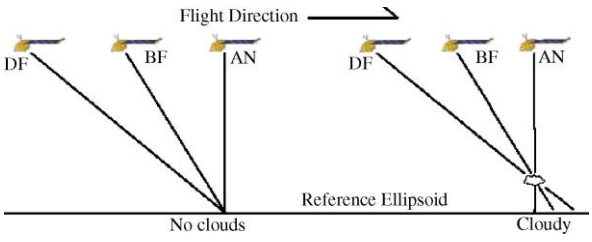

图2.地表特征和云层与参考椭圆体的登记。请注意，9台MISR相机中只有3台是准确的，表面物体被登记到同一位置，而云层被登记到不同的位置。

不同MISR相机的椭圆体位置。L2TC算法与不同相机图像中的同一物体相匹配，允许通过简单的三角关系检索物体的高度和水平速度。SDCM是通过比较检索到的物体高度和已知的地形高度得出的，超过地形高度约650米的物体被归类为云。

ASCM的基础是带差角标（BDAS），它描述了两个太阳光谱反射率之差与视角之间的关系（Di Girolamo和Davies 1994；Diner等人1999b）。根据空气分子（瑞利）散射对太阳光反射的相对贡献，太阳光在到达MISR之前经过地球大气层的时间越长，BDAS越大。因为从云层反射的辐射比从地球表面反射的辐射路径短得多，所以小的BDAS表示云层。在Arctic的冰雪覆盖的表面上，来自Df相机的蓝色辐射和来自Cf相机的近红外辐射的大小之间的差异被阈值化，以区分云和表面。

这两种算法在极地地区都有困难，特别是在北极地区白天经常出现的低云。SDCM不能探测到低云，因为低云和地表之间的差距往往小于MISR的高度检索精度。ASCM善于探测高云和薄云，但很难探测到地形上的低云，因为Rayleigh散射对BDAS的贡献随着云顶高度的降低而迅速增加（Di Girolamo和Davies 1994）。请注意，SDCM和ASCM都使用云的特征来寻找云。我们在第3节提出的方法，试图通过搜索地表像素而不是云层像素来克服SDCM和ASCM的这些缺点。

2.2 数据

本研究中使用的数据是从北极、格陵兰岛北部和巴芬湾上空的10个MISR轨道26位收集的。如前所述，在同一路径上的两个结果轨道之间的重复时间是16天，所以这10个轨道从2002年4月28日到9月19日（北极地区的一个白昼季节）大约跨越144天。我们选择路径26进行研究，因为它的表面特征非常丰富，其中包括北冰洋的永久海冰、格陵兰岛被雪覆盖和无雪的沿海山脉、永久冰川雪和冰，以及在研究期间在巴芬湾融化的海冰。

本研究包括每个轨道上的六个数据单元（MISR第11-13、14-16、17-19、20-22、23-25和26-28块）。60个数据单元中的3个被排除在本研究之外，因为在夏季海冰融化后，表面是开放的水面，而MISR的操作算法可以很好地检测水面上的云层。因此，调查的数据包括57个数据单元，有7,114,248个1.1公里分辨率的像素，每个像素有36个辐射测量值。我们使用275米的红色辐射测量值来建立我们的一些特征，所以数据集的实际规模甚至更大。我们的研究集中在随着时间的推移进行重新访问，以便专家熟悉地表特征，这极大地帮助了专家的标注过程。

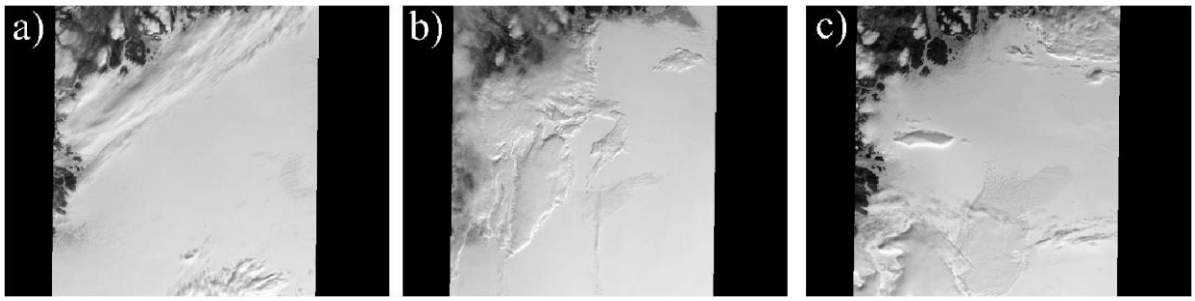

图3.MISR An相机在路径26的20-22区块上连续三个轨道（即13257、13490和13723）收集的数据。

图3显示了在三个连续的轨道上从MISR第20-22块收集的三个数据单元的图像（即13257、13490和13723）。这些图像显示，冰雪覆盖的表面和云一样亮，甚至比云更亮，这与适用于低纬度地区的传统假设相矛盾，即云比表面更亮。因此，基于这一假设的传统云层探测算法不能在极地地区应用。此外，不同斑块图像之间的差异并不是云层与地表的必然结果，因为整个数据单元可能是完全被云层覆盖或完全清晰。

为了评估我们提出的方法和现有的MISR操作算法的性能，其中一位作者对数据进行了手工标注。由于高质量的地面测量数据在极地地区很少，因此专家对晴天和阴天场景的标注是目前产生评估数据的最佳方法，用于评估自动极地云层探测算法。在专家确定了一块图像像素的标签（即晴天或多云）后，喷气推进实验室的MISR科学团队开发的工具，称为 "mis- rdump "和 "misrlearn"（Dominic Mazzoni，个人通讯），随后被用来标记MISR天底相机图像中的像素为晴天或多云。例如，图3所示数据的标签被绘制在图4中，白色代表阴天，灰色代表晴天。专家标签只给那些专家根据他的测量知识非常有信心的像素，而更多的模糊区域则不加标签（黑色像素）。由于这种保守的标签方案，专家标签覆盖了总有效像素的71.5%（5,086,002）。我们使用这些标签来评估不同的云检测算法的性能。

3. 方法

我们为Arctic提出了两种新的云层检测算法，这些算法对于大规模MISR数据的操作处理来说是自动化的、快速的。这两种算法都是基于三个生理上有用的特征，用于识别表面像素，从而通过排除法获得多云的像素。每个像素都被视为独立于其他像素的，因为大多数空间模型的计算成本太高，无法在业务数据处理中实施。在我们的算法中隐含地考虑了地表和云层像素的空间平滑性，因为我们的三个特征中有两个是基于275米分辨率下的局部像素斑块。准确地说，我们解决这个北极云层探测问题的策略概括为三个步骤。

- *步骤1*.基于EDA和领域知识构建三个特征（第3.1节）。
- *第2步*。通过在每个特征上设置阈值（固定或数据自适应），建立ELCM云检测算法。
  并对每个数据单元应用ELCM，产生第一个云检测产品（第3.2节）。
- *步骤3*.通过训练预测部分多云数据单元的多云概率，即第二种多云检测产品。
  在由ELCM算法产生的标签上采用Fisher的QDA（第3.3节）。

回顾一下，专家标签并不能用于在线数据处理。在步骤2中，只有对第一天收集的数据单元，我们才用离线的专家标签为这三个特征设定最佳阈值。对于同一区块范围内的后续数据单元，阈值是通过将正在处理的数据单元的聚类信息与从前次访问中学到的阈值相结合来确定的。

图4.MISR轨道（a）13257，（b）13490，和（c）13723的第20-22块的专家标签。白色代表高置信度的阴天；灰色代表高置信度的晴天；黑色代表未标记的像素。

3.1 构建三个物理特征

MISR北极云层探测提出了一个巨大的数据分析问题。对于每个像素，有36个维度（4个波长，9个角度）。在每个数据单元中，在1.1公里的分辨率下有196,608（=384行×512列）像素。在我们的研究中，只有MISR的红色辐射数据被用于构建------。
摄取特征，原因有二。首先，基于探索性的数据分析和液态和冰态水颗粒的散射特性，所有四个波段在冰、雪和云上都有类似的反射图谱（Clothiaux, Barker, and Ko-rolev 2005）。第二，只有红色辐射有275米的空间分辨率。
分辨率，是MISR的最高空间分辨率。
在这项研究中，我们使用的辐射测量值是亲...
与底层表面相连接（即表面注册数据）。
在MISR的运行处理过程中，MISR的红色辐射测量值以两种不同的方式登记到参考椭圆体上。
有不同的方法。一种方式是图2所示的参考椭球体登记，即辐射测量值直接登记到参考椭球体上，而忽略了地形。在第二个地形投影登记过程中（图5），辐射测量值被投影（在陆地上）到雨季表面，然后给定参考椭球体坐标

的地形地貌。表面注册的数据指的是椭圆体的
在海洋上的预测测量和地形上的预测测量。
对陆地表面的测量。
在所有的数据单元中，我们调查了大量特征的分布情况，包括Angu-的线性组合。
我们发现，在不同的角度和波长之间的相关性、辐射的非线性转换、云层的空间模式和反射表面的光滑度。结合广泛的探索性数据分析和具体的领域知识，如冰雪表面比云层有更多的前向散射，更多的各向同性散射，我们发现了三个物理上有用的特征，可以区分地表像素和云层。

让$I^k_{m,\ell}$表示275米分辨率红光的（*米*，*英镑*）像素。对*第k个*MISR相机角度的辐射测量，其中k从1到9，代表MISR相机Df到Da。

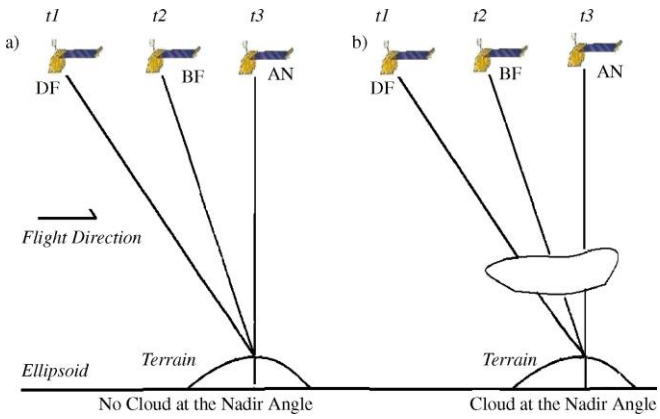

图5：(a)在没有云的情况下，MISR辐射测量值与地形的登记。(b)
在有云的情况下，MISR辐射测量值与地形的登记。请注意，来

*3.1.1 第一个特点。CORR。*第一个特征，CORR，来自Shi, Yu, and Braverman (2002)，是不同视角下辐射测量的平均线性相关性。我们在2.2公里的空间分辨率下定义CORR，但是
使用275米数据，间距为1.1公里。一个数据单元中的每个1.1公里间距、2.2公里分辨率的相关*系数*LC_{ij}（$i = 1,\ldots, 384$; $j = 1,\ldots, 512$）是由275米分辨率的8×8组像素构成的。利用这些8×8 275米分辨率的像素，我们用以下方式表示k_1和k_2 MISR观测方向（$k_1 \neq k_2$和$1 \le k_1, k_2 \le 9$）之间的线性相关系数LC_{ij}

$$LC^{k1,k2}_{ij} = \frac{\sum_{m=4i-5}^{4i+2}\sum_{\ell=4j-5}^{4j+2}(I^{k1}_{m,\ell} - \overline{I^{k1}_{i,j}})(I^{k2}_{m,\ell} - \overline{I^{k2}_{i,j}})}{\sigma^{k1}_{i,j}\sigma^{k2}_{i,j}}, \quad (1)$$

其中，$I^k_{m,\ell}$是位置(m, ℓ)的275米分辨率辐射测量值，$\overline{I^k_{i,j}}$和σ^k_{ij}是与位置(i, j)相关的*第k*个角度的辐射测量值的平均值和标准偏差。
平均数是算术平均数，而标准差是算术平均数。
浓度由以下公式给出

$$c^k_{i,j} = \frac{1}{64-1}\sum_{m=4i-5,\ell=4j-5}^{4i+2,4j+2}(I^k_{m,\ell} - \overline{I^k_{i,j}})^2。 \quad (2)$$

我们期望在无云区或低云区有高的相关性。
因为每个MISR相机测量的辐射都是从同一个表面散射出来的（图5a）。云层，特别是高云层，预计辐射测量的相关度较低（图5b）。然而，由于云层的运动，云层的高相关性可能在极少数情况下出现。因此，我们将第一个特征定义为从MISR
Af/An和Bf/An相机得到的相关系数的平均值。
对。

$$corr_{ij} = (LC^{4,5}_{ij} + LC^{3,5}_{ij})/2,$$

自云层不同部分的测量值被映射到地形上的同一位置。

因为在实践中，对于高海拔的云层，这个平均值总是相对较低。

CORR的高值表明，要么是晴朗（无云）的条件，要么是在底层表面的同一位置存在低高度的云。但如果我们对高 CORR 的像素宣布为晴天，对低 CORR 的像素宣布为阴天，那么我们就会产生两种类型的错误。平滑无云的地形表面可以被归类为多云（低信噪比导致低关联度），而低高度的云可以被归类为清晰。为了避免这些错误，我们还需要两个特征。

*3.1.2 第二个特点。SD。*为了识别光滑的表面，MISR An（k = 5）相机红辐射测量组内的标准偏差[公式（2）]是有用的。

$$SD_{ij} = \sigma^{5}_{i,j}$$

。

对于从光滑表面发出的辐射来说，它的数值很小，在这种情况下，不同的MISR观测方向之间的相关性被仪器噪声所支配。

*3.1.3 第三个特点。NDAI。*第三个特征是由Nolin、Fetterer和Scambos（2002）开发的NDAI。

$$NDAI_{ij} = \frac{\overline{I^1_{ij}} - \overline{I^5_{i,j}}}{\overline{I^1_{ij}} + \overline{I^5_{i,j}}}, \quad (3)$$

其中平均辐射测量值是在4×4组275米分辨率的红色辐射测量值上（因此空间分辨率为1.1公里）。

这一特点是由于在可见光波长下，来自冰雪覆盖的表面的各向异性比来自低海拔云层的更多（Stephens, Cambell, and Vonder Haar 1981）。在我们目前的实施中，如(3)所述，MISR Df-相机的测量结果与An相机的测量结果进行了比较。在北极地区的云层上，Df相机收集的辐射要比An相机收集的辐射高得多；因此，较大的NDAI值表明有云层存在。

我们的调查表明，在CORR、SD_{An}和NDAI所跨越的空间中，晴朗和多云的像素分离得很好。此外，这些特征对于建立一个有效的云层检测算法来说还有一个理想的特性。对于来自不同地理位置或不同时间的数据，来自清晰表面的特征分布是差不多的。我们在第3.2节中说明了在我们的组合聚类和分类算法中特征的这种稳定性的重要性。不幸的是，MISR辐射测量值本身并不具有这些"可分离性"和"稳定性"的特性（详见本文的长篇文章。见*www.stat.berkeley.edu/~binyu/publications.html*）。

3.2 增强型线性相关匹配算法

如第1节所述，对于MISR数据的操作处理，没有专家标签来训练每个数据单元的检测器。因此，检测器必须在没有专家标签的数据单元或以前收集的数据单元上进行训练。此外，聚类算法的结果并不总是重新反映出晴天和阴天之间的区别，特别是当数据单元完全被云覆盖或完全晴天时。

在这一节中，我们提出了一种操作算法，该算法将当前数据单元的聚类信息与从以前的数据单元学到的检测因子结合起来。当前数据单元上的聚类结果的质量取决于特征空间的"可分离性"属性。借用以前的数据单元开发的检测器，要求事先收集的清晰和浑浊像素在特征空间的分布与当前数据单元的分布相似。前面描述的"稳定性"和"可分离性"是我们物理特征的两个重要属性，对我们的方法的成功至关重要。不幸的是，在辐射测量的空间中，由于不平坦的地形和太阳辐射大小的太阳光照方向的变化，晴天和阴天的像素既不容易分离，也不可能在不同的时间和空间中稳定。因此，我们的特征开发对于我们方法的成功既是必要的，也是至关重要的。

鉴于上述三个特征的两个理想特性，一个自然的、计算效率高的检测方法是通过对特征进行阈值处理来标记像素。阈值要么是固定值，要么是由数据适应性选择的值。方法。我们称这种方法为ELCM，因为在Shi等人（2002）的线性相关匹配分类（LCMC）中增加了两个额外的特征。我们现在详细描述这个过程。

3.2.1 ELCM算法。

一个1.1公里×1.1公里的像素在以下情况下被标记为清晰的

- $SD_{An} < 阈值_{SD}$ **或**
- $CORR > 阈值_{CORR}$ **，** $NDAI < 阈值_{NDAI}$。

当上述测试失败时，该像素被标记为多云。

ELCM算法的原理直接来自于我们对三个特征的选择。在晴朗的天空下，冰雪覆盖的表面要么非常光滑，具有极小的SD_{An}，要么在不同视线方向之间具有高度的相关性和相对较低的向前散射。云层很少，如果有的话，也是非常光滑（在不同视角之间产生低的相关性）和相对较弱的向前散射。在调查了来自不同轨道的专家标记的晴天和阴天像素的分布情况，并比较了从不同阈值到专家标记的结果后，我们得出结论，CORR和SD_{An}的阈值，在所有的数据单元中都是稳定的，而且很好用，我们将它们设定为经验确定的固定值：

$$阈值_{CORR} = .75, \quad 阈值_{SD} = 2.0。$$

不幸的是，NDAI的适当阈值因数据单位不同而不同，因此必须适应性地学习$阈值_{NDAI}$。

3.2.2 阈值的选择$_{NDAI}$。我们在当前的数据单元上使用聚类算法，以及从相同位置的以前的数据单元学到的阈值来设置$阈值_{NDAI}$。对于第一个轨道上的每个数据单元，我们从[0, 1]中对$阈值_{NDAI}$进行网格搜索，其步骤为1×10^{-5}，以确定导致最小分类的数值。相对于专家标签的计算误差。对于每个数据单元，在后来的轨道，我们通过*k-means*算法（Dempster, Laird, and Rubin 1977）初始化的EM算法将两个高斯分布的一维混合模型拟合到观察到的NDAI中。两种成分之间的倾角通过以下方式找到
在双分量高斯的两个拟合均值之间进行相同步长为1×10
的数值网格搜索$^{-5}$。
混合分布。如果存在凹点，则被用作阈值，它通常落在经验确定的范围内，即
.08-
.40.否则，就使用前一个（16天之前）数据单元的阈值。在拟合NDAI分布时，我们修剪掉上下2.5%的尾部，以避免两个尾部的离群值和极端值，从而提高拟合方案的稳健性。出于同样的稳健性考虑，我们使用了拟合分布中的倾角，而不是贝叶斯规则的截止点。

图6显示了一个NDAI分布和与之相适应的混合模型的例子。这些结果是基于13490号轨道的MISR20-
22块。对于这个数据单元，拟合分布中的凹陷相当明显，发生在0.215的值上，落在0.08到0.40的预期范围内，而且

不需要前次的调整的阈值。使用 *阈值*$_{CORR}$ = .75，*阈值*$_{SD}$ = 2.0，*阈值*$_{NDAI}$ =

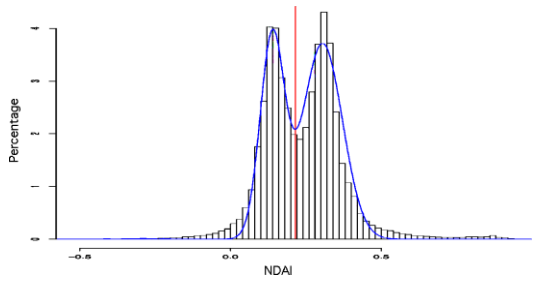

图6.13490号轨道20-22区块的NDAI直方图（竖条），以及两个一维高斯函数混合模型对它的拟合（蓝线）。垂直红线表示从这个直方图中得出的NDAI阈值。

.215对数据单元中的像素进行分类，我们获得了与专家标签96.16%的一致率。13257号和13723号轨道的MISR区块20-22的准确率（图7）为87.53%和94.36%。

当该算法在网上实施时，可能会进一步改善*阈值*$_{\text{NDAI}}$的选择。由于本案例研究中的计算限制，我们只调查了从路径26收集的数据，这导致两次访问之间有16天的时间滞后。对于全球数据来说，在极地地区，两个连续的MISR访问之间的时间滞后实际上比16天短得多，因为MISR路径在高纬度的空间上有很大的重叠。因此，从上一次访问中学到的阈值应该比16天前学到的阈值表现得更好，我们期望ELCM算法的表现甚至更好。

3.3 通过在ELCM上训练QDA进行概率预测

ELCM算法旨在为每个具有有效辐射测量的像素提供清晰或浑浊的检测标签。由于该算法产生的阈值并不完美，其标签也不完美。此外，如果一个像素是部分阴天，并不总是能够准确地将其归类为晴天或多云，因为在云层边界附近总是不同程度地发生。因此，报告多云的概率是可取的，比只提供一个二进制的晴天与多云的标签更有参考价值。我们使用从ELCM标签中训练出来的Fisher's QDA来提供对多云的概率或可信度的估计。

如果ELCM的结果显示，整个数据单元几乎完全清晰或浑浊，98%的标签属于同一类别，那么我们只报告标签。否则，ELCM的标签将被作为作为输入数据来训练QDA的三个特征。在本研究的57个数据单元中，32个是部分阴天，25个是完全晴天或阴天。

在两类分类问题中，QDA将每个类的密度建模为多变量高斯分布N(μ_s , $:E_s$)，其中s = 1，2表示类标签。让π_s是先验概率
后验概率P（$x\in$类$s|x$）为
由贝叶斯法则给出。参数π_s , μ_s , 和 $:E_s$ 是es-
经验（即训练数据）的等级比例、平均值和协方差矩阵，然后代入上述两个方程式。估计的后验概率P（$x\in$类$s|x$）作为一个像素的多云性预测概率。

与其他只对条件概率$P(s|x)$建模的方法相比，如逻辑回归，QDA在计算上更有效率，对训练标签的错误也更稳健，因为它对联合分布$P(s,$
$x)$建模。QDA的计算只涉及到对P(s, x)的向量进行估计。平均值和协方差矩阵，这比逻辑回归的迭代加权再拟合要有效得多。在我们的设置中，X的均值和协方差是由ELCM阈值算法产生的标签估计的。这些标签包含由于阈值选择而产生的错误。然而，如果标签中的错误在整个训练样本中只占很小的比例，那么标签中的错误就不会对均值和协方差的估计产生很大影响。

Logistic回归只对$P(s|X)$的条件分布进行建模。当它根据ELCM算法产生的标签进行拟合时，逻辑回归对标签保持忠诚，因此受ELCM再结果中错误标签的影响更大。简单地说，在这个问题上，QDA比逻辑回归的效果更好。
的回归。这一现象也是最近对半监督学习兴趣大增的原因（Zhu, Ghahramani, and Lafferty 2003; Zhou, Schölkopf, and Hofmann
2005），其目的是考虑到预测者的分布，或特征分布。我们还测试了基于最小协方差行列式估计的QDA的稳健版本（参见Rousseeuw 1985；Croux和Haesbroeck 1999），但它并没有改善估计值或标签结果。因此，我们使用ELCM-QDA来提供概率标签。

4. 结果

为了评估ELCM和ELCM-QDA算法，以及MISR操作SDCM和ASCM算法的性能，我们现在将它们的结果与专家拉贝尔进行比较。除了我们提出的操作算法外，还有

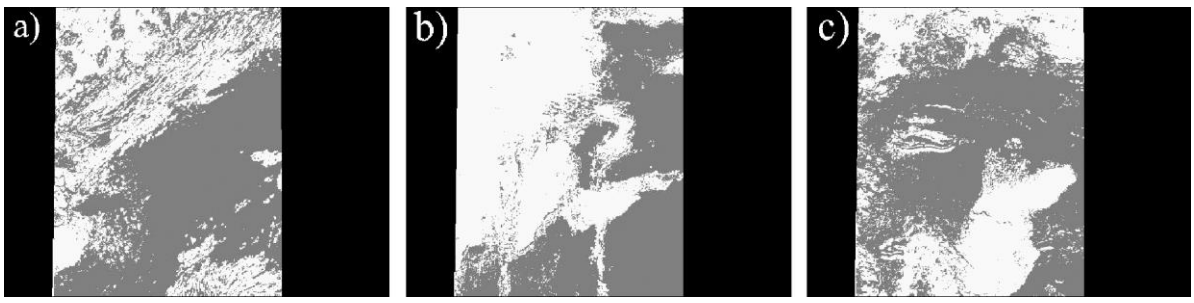

图7.MISR轨道（a）13257、（b）13490和（c）13723的第20-22块的ELCM算法结果；这些数据单元与图3和图4中的相同。白色代表归类为多云的像素；灰色代表归类为清晰的像素。

在不使用专家标签的情况下（除了第一个轨道），我们还测试了一个简单的离线SVM（Vapnik 1995）分类器，该分类器是在随机抽样的专家标签上训练的，以说明我们特征选择和聚类/分类框架的价值。我们讨论了ELCM算法和专家标签之间一致率最低的数据单元的结果，以提供对ELCM在某些地理位置上失败的原因的洞察。除了比较所有可用的操作算法的结果，我们还说明了三个MISR特征的可分离性和稳定性，并解释了为什么我们的特征选择是必要的，而且效果很好。

4.1 ELCM、ELCM-QDA、ASCM、SDCM和SVM的比较

在研究的57个数据单元的像素中，有5,086,002个像素是由专家标注的；这些像素占所有像素的71.5%。此外，如前所述，一个使用MISR红色辐射的离线高斯克尔SVM也在前20个数据单元中随机选择的20,000个专家标签（每个数据单元1,000个）上进行训练，并在其余37个数据单元中进行测试。这个SVM分类器的准确性表明，当以一种简单的方式应用于这个问题时，最先进的分类器表现得很好。SVM的训练和测试程序遵循标准的10倍交叉验证方法。除了准确率，我们还报告了算法的覆盖率，定义为提供标签的像素的百分比。根据设计，ELCM、ELCM-QDA和SVM提供100%的覆盖率，而ASCM和SDCM在57个数据单元中的覆盖率为70.12%和26.64%。

表1中的结果显示，在500万个测试像素中，ELCM算法的一致率为91.80%，比MISR ASCM（83.23%）算法高8.57%，比SDCM（80.00%）算法高11.90%。这从科学和统计学的角度看都是一个重大的改进。当使用专家标签进行训练时，离线SVM的同意率为80.99%，远远低于ELCM算法（但与SDCM或ASCM算法相当）。离线SVM的低准确率可以解释为，晴天和阴天像素的分布都取决于数据采集的时间和地点，但一个单一的分类器不能适应时间或地点。ELCM算法是以一种依赖数据的方式建立的，并能适应分布的变化。

相对于ELCM算法，ELCM-QDA算法并没有提高与专家标签的总体同意率。

表1.ELCM、MISR操作SDCM和ASCM以及离线SVM算法相对于专家标签和覆盖率的一致（准确）率

	ELCM	SDCM	证券交易委员会(ASCM)	离线SVM
与专家的协议	91.80%	80.00%	83.23%	80.99%
覆盖范围	100%	26.64%	70.12%	100%

然而，ELCM-QDA算法超越了ELCM的阴天与晴天的双重标签，它提供了概率labels。图8显示了在图3、4和7中讨论的三个数据单元的ELCM-QDA结果。ELCM-QDA算法表现合理的一个指标是在云层边界（图4的灰色区域）经常出现0.5左右的概率（图8的绿色区域）。我们期待这一结果，因为边界像素处于晴朗和多云区域之间，这与我们使用概率标签的动机一致。

图9(a)显示了57个单独数据单元的ELCM算法和专家标签一致率的柱状图。在所有的数据单元中，在大多数情况下，ELCM算法与专家标签的一致率达到或远远超过90%。有几个数据单元的吻合率低至70%，我们现在对其中一个数据单元进行分析，以确定ELCM算法或专家标签中是否存在任何系统性偏差。

由13956号轨道17-19号区块组成的数据单元，其ELCM算法结果与专家标签之间的一致率最低（71%）。我们发现，两者之间的分歧主要发生在专家标注的清晰区域上，而这些区域通过进一步的调查被确认为是准确的。ELCM算法再结果中出现错误的主要原因是，即使没有云，CORR也很低。在这种情况下，低相关性是由于不良的地形数据注册的结果，这通常发生在急剧的海拔变化（例如格陵兰岛的海岸线）。这导致ELCM算法的结果在地形粗糙的地区出现系统误差。目前，我们还没有解决这个问题的办法，因为这需要更新MISR的地形高度数据库，而这在目前还没有计划。

4.2 需要进行特征选择和自适应阈值处理

随机抽样的专家标签被用作训练数据，以证明我们三个特征的可分离性和稳定性。

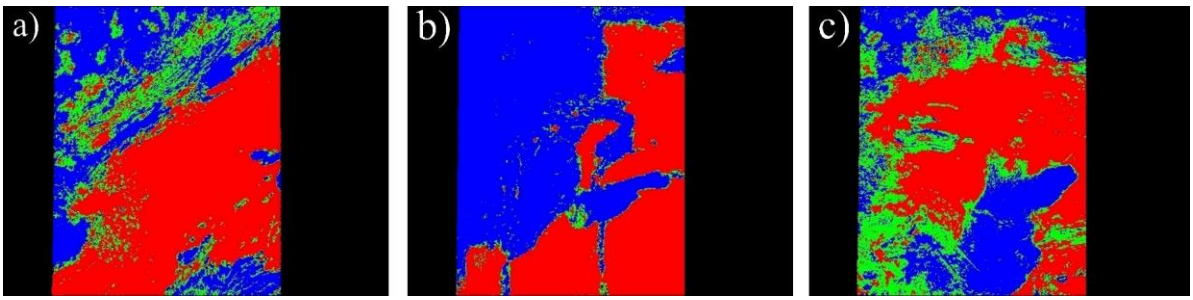

图8.MISR轨道（a）13257、（b）13490和（c）13723的第20-22块的ELCM-

QDA结果，这与图3、4和7所示的数据单元相同。红色代表*P(*cloudy|x*)*< .2的像素；绿色代表.2 < *P(*cloudy|x*)* < .8；蓝色代表*P(*cloudy|x*)* > .8.

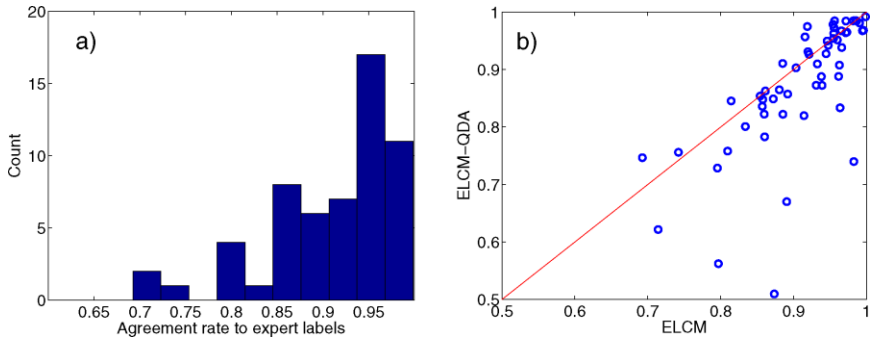

图9：（a）研究中57个单独数据单元的ELCM算法和专家标签的同意率直方图。(b) 研究中32个部分云数据单元的ELCM-QDA算法与ELCM算法与专家标签的一致率。

相对于辐射测量。为了比较可分离性，在一个数据单元的一半专家标签上训练QDA、逻辑回归和带有*L1*惩罚的逻辑回归，并在同一数据单元的另一半上进行测试。结果表明，基于特征的检测器总是优于那些只基于辐射测量的检测器。为了比较稳定性，我们将用一个轨道（即13257）的专家标签训练的QDA和Logit应用于随后的一个轨道（即13490），发现基于特征的检测器优于基于辐射的检测器（详见本文的长版，网址为 *www.stat.berkeley.edu/~binyu/publications.html*）。

适应性地寻找NDAI的阈值的选择是超...
下面的实验证明了这一点。用我们的特征作为预支配因素，在前一次访问（轨道13257）上训练的QDA对当前（轨道13490）数据单元的分类比在当前数据单元上训练的QDA更差，准确率为92.6%，准确率为95.9%。如图10所示，检测器性能的下降主要是由于不同访问之间NDAI分布的差异。

5. 总结、结论和影响

本文介绍了一项统计研究，这只是为准确描述日光下北极地区的云层特性而必须进行的众多研究中的第一个。我们已经证明了三个物理特征--
来自不同MISR观测方向的辐射测量值的线性相关（CORR），MISR天底红辐射测量值在一个小区域内的标准偏差（SD_{An}），以及一个标准化的差异角指数（NDAI）--
包含足够的信息，将云与冰雪覆盖的表面分开。基于这三个特征的ELCM算法比现有的用于北极地区云层探测的MISR操作算法更加准确，并提供了更好的空间覆盖。ELCM算法结合了分类和聚类框架，使其适用于实时、操作性的MISR数据处理。计算速度足够快，能够以在线方式处理大量的MISR数据流。ELCM算法的结果可用于训练QDA，为部分云雾场景提供概率标签。

这项工作在两个方面对统计学具有重要意义，这超出了统计学方法的技术开发和实施。在过去的十年里，支持天气和气候研究的地球科学数据大量涌现。将这些数据应用于当前的科学问题，如飓风预测和气候变化，需要在个案的基础上对数据应用适当的统计方法。因此，如果统计学家选择这样做，他们在这些数据的分析中可以发挥重要作用（Yu
2006）。在这项关于白天极地地区云层检测的研究中，我们发现三个特征提供了足够的可分离性和稳定性，可以将透明（无云）地区与云层分开，一个不比QDA更复杂的分类器所提供的性能可以与更复杂的分类器（如SVM）相媲美。在过去，统计学家经常以一种模式工作，即他们在事后与科学界的同事一起工作，开发事后分析方法------
。

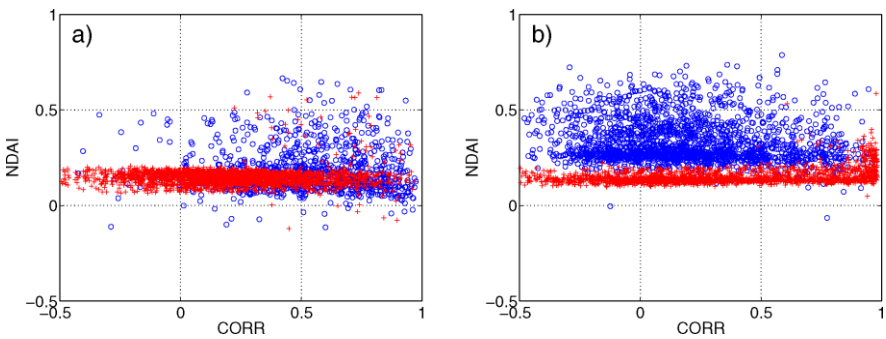

图10.(a) 13257和(b) 13490轨道的MISR第20-22块的NDAI与CORR的散点图。蓝色圆圈代表专家标示的多云像素；红色圆圈代表专家标示的清晰像素。

统计学。然而，在这里，统计学家直接参与了数据处理的具体工作。只有在大气科学家、统计学家和喷气推进实验室的MISR科学和仪器团队之间有时具有挑战性的合作之后，才能获得成功。

这项研究的第二个重要方面是，它展示了统计思维的力量，以及统计学为现代科学问题提供解决方案的能力。当然，这在统计学界是没有争议的，但如果从其他没有受过训练的人（如美国国家航空航天局的管理人员）的角度来看，我们这门学科的贡献就不那么明显。许多人不明白，除了精通统计学的计算机和大气层科学家所提供的，统计学还有什么贡献。这项工作表明，通过对已证实的统计方法的仔细组合和特定应用的修改，人们可以为明显复杂的科学问题制定有效、创新的解决方案。如果没有对统计学原理的基本了解，就不可能制定这样的策略，如果没有沉浸于特定领域的科学基础和问题的意愿，也不可能制定这样的策略。

从科学的角度来看，毫无疑问，在极地场景中探测云层的能力是一个非常重要和有趣的话题。2007年迎来了一项重要的国际倡议，以研究和更好地了解地球的两极进程--2007年国际极地年（IPY）。上一个国际极地年，即1957-58年，有来自67个国家的大约8万名科学家参与。50年后，目前的国际极地年将包括来自卫星传感器的信息，包括MISR。来自MISR的经证实的云层掩码，结合来自Terra卫星上的MODIS的掩码，将使一系列经证实的云层和辐射研究成为可能（Shi, Clothiaux, Yu, Braver-man, and Groff 2007）。这种云层研究的最终目的是提高对通过大气层的可见和红外辐射流的理解，因此我们可以开始区分云层对北极气候变化的反应以及它们对气候的反馈。更可靠的极地云层特性也将通过改进模型云层物理学转化为更准确的全球气候模型模拟。这些研究最终将使科学界能够研究云层特性的变化如何加强或改善大气中二氧化碳浓度增加给北极带来的任何初步变化。

[2006年8月收到。2007年4月修订。]

参考文献

Clothiaux, E. E., Barker, H. W., and Korolev, A. V. (2005), "Observing Clouds and Their Optical Properties, " in *3D Radiative Transfer in Cloudy Atmospheres*, eds.A. Marshak 和 A. B. Davis, 柏林。Springer，第93-150页。

Croux, C., and Haesboeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161-190.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Di Girolamo, L., and Davies, R. (1994), "A Band-Differenced Angular Signature Technique for Cirrus Cloud Detection," *IEEE Transactions on Geo-Science and Remote Sensing*, 32, 890-896.

Diner, D. J., Asner, G. P., Davies, R., Knyazikhin, Y., Muller, J.-P., Nolin, A. W., Pinty, B., Schaaf, C. B., and Stroeve, J. (1999a), "New Directions in Earth Observing Scientific Applications of Mulitangle Remote Sensing, " *Bulletin of American Meteorological Society*, 80, 2209-2228。

Diner, D. J., Braswell, B. H., Davies, R., Gobron, N., Hu, J. N., Jin, Y. F. , Kahn, R. A., Knyazikhin, Y., Loeb, N., Muller, J.-P., Nolin, A. W., Pinty, B., Schaaf, C. B., Seiz, G., and Stroeve, J. (2005), "The Value of Multiangle Measurements for Retrieving Structurally and Radiatively Consistent Proper- ties of Clouds, Aerosols, and Surfaces, " *Remote Sensing of Environment*, 97, 495-518。

Diner, D. J., Davies, R., Di Girolamo, L., Horvath, A. Moroney, C., Muller, J.-P., Paradise, S. R., Wenkert, D., and Zong, J. (1999b), "MISR 2级云检测和分类算法的理论基础，" JPL技术文件D-11399, rev.D, Jet Propulsion Laboratory, Pasadena, CA.

Fisher, A. R. (1925), *Statistical Methods for Statistical Workers*, Edinburgh:Oliver and Boyd.

Francis, J. A., Hunter, E., Key, J. R., and Wang, X. (2005), "Clues to Variabil- ity in Arctic Minimum Sea Ice Extent, " *Geophysical Research Letters*, 32, L21501, doi:10.1029/2005GL024376.

Giorgi, F., and Bi, X. (2005), "Update Regional Precipitation and Temperature Changes for the 21st Century From Ensembles of Re-cent AOGCM Simulations," *Geophysical Research Letters*, 32, L21715, doi:10.1029/2005GL024288.

Kato, T., Tang, Y., Gu, S., Hirota, M., Du, M., Li, Y., and Zhao, X. (2006), "Temperature and Biomass Influences on Interannual Changes in CO_2 Exchange in an Alpine Meadow on Qinghai-Tibetan Plateau," *Global Change Biology*, 12, 1285-1298.

Mardia, K. V., Kent, J. T., and Bibby, S. M. (1979), *Multivariate Analysis*, New York:学术出版社。

Nolin, W. A., Fetterer, M. F., and Scambos, A. T. (2002), "海冰和冰盖的表面粗糙度特征。使用MISR数据的案例研究"，*IEEE地球科学和遥感论文*，40，1605-1615。

Rousseeuw, J. P. (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications*, eds.W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: B. Reidel.Reidel, pp.283-297.

Shi, T., Yu, B., and Braverman, A. J. (2002), "MISR Cloud Detection Over Ice/Snow Using Linear Correlation Matching," Technical Report 630, University of California Berkeley, Dept. of Statistics。

Shi, T., Clothiaux, E. E., Yu, B., Braverman, A. J., and Groff, D. N. (2007), "Detection of Daytime Arctic Clouds Using MISR and MODIS Data, " *Remote Sensing of Environment*, 107, 172-184.

Stephens, G. L., Campbell, G. G., and Vonder Haar, H. T. (1981), "Earth Radi- ation Budgets," *Journal of Geophysical Research*, 86, 9739-9760.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, New York:Springer.

Yu, B. (2006), "拥抱信息技术时代的统计挑战"， 技术报告706, 加州大学伯克利分校统计系。

Zhou, D., Schölkopf, B., and Hofmann, T. (2005), "Semi-Supervised Learning on Directed Graphs," in *Advances in Neural Information Processing Systems*, eds.L. K. Saul, Y. Weiss, and L. Bottou, Cambridge, MA:麻省理工学院出版社，第17页。

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003), "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, " *ICML*, 20.