A 10 m resolution global lake dataset based on Sentinel-2 MSI data and deep learning

Beihui Hu1, and Lian Feng1

2 Materials and Methods

在本节中，我们介绍了所使用的数据及数据集制作的技术流程（Fig. 1）。我们利用Sentinel-2 L2A地表反射率产品进行全球湖泊边界提取，并结合多个辅助数据集改进预测结果。湖泊提取方法如下：1) 数据预处理，在Google Earth Engine 平台上对Sentinel-2数据进行处理，计算NDWI指数，对各波段进行多年逐像素平均，将所需数据下载至本地计算机。2) 模型训练，利用在全球范围选取的训练数据集，对模型进行精度评价。3) 对全球湖泊进行预测。4)后处理操作，结合多个辅助数据集删除错分的湖泊，补全缺漏的湖泊，获得最终的数据集GLAKESplus。

Fig. 1. Flowchart for developing the GLAKESplus dataset.

2.1 Data sources

2.1.1 Sentinel-2

在GLAKESplus制图中使用了Sentinel-2数据。Sentinel-2是欧洲航天局( ESA )哥白尼计划的高分辨率多光谱成像任务，由2015年发射的Sentinel-2A 及2017年发射的Sentinel-2B组成，重访周期为2~5天。其B2(蓝光,0.46~0.52 μm)、B3(绿光,0.55~0.58 μm)、B4(红光,0.64~0.67 μm)及B8 (近红外,0.78~0.90 μm)波段空间分辨率为10米，B11(短波红外波段,1.57~1.65 μm)为20米。我们于GEE平台获得研究时段（2017年3月28日至2022年4月10日）内云像素比例低于60%的Sentinel-2 L2A产品。

利用可见光影像进行湖泊边界提取需要解决云及云阴影、冰雪、泥沙、藻华等因素的干扰，因此需要进行一系列预处理操作。本文利用哨兵云概率产品进行云掩膜，利用L2A的场景分类图层进行冰雪掩膜。哨兵云概率(S2 Cloud Probability, S2Cloudless) 产品是使用 sentinel2-cloud-detector 库创建的一个云概率产品，提供了每幅影像的单个像素的云概率。本文去除了云概率大于50%的像素，并将近红外波段的暗像元与云投影取交集，从而去除了云阴影。Sentinel-2 L2A产品的场景分类图层 (Scene Classification Layer, SCL) ，对云、云阴影、冰雪、水体进行了初步的分类，利用SCL波段能去除Sentinel-2影像中大部分冰雪信号。随后，本文计算了NDWI水体指数，并对研究时段内的影像数据进行逐像素平均操作。NDWI指数 (McFEETERS, 1996)是一种常用的水体指数，利用绿光与近红外波段的波段组合，能够增强水体信息，在Sentinel-2中，由计算。湖泊的面积在一年内动态变化，对影像进行平均，能够获得湖泊的平均水面积，并削弱单幅影像的云残留、泥沙、藻华等信号的干扰。最终，本文将平均的可见光波段(B2, B3, B4)、短波红外波段（B11，重采样至10m）及NDWI波段按照1°×1°格网下载至本地计算机，作为后续湖泊预测的底图。尽管短波红外波段分辨率仅20m，但该波段有助于模型区分建筑物与水体（NDWI指数无法很好地区分二者）。

2.1.2 Other lake dataset

为了获得覆盖更全面的全球湖泊数据集， GLAKES和PLD湖泊数据集被用于补全缺漏的湖泊。GLAKES是Pi等(Pi et al., 2022)基于GSW水概率底图与深度学习方法制作的全球湖泊数据集，涵盖了全球约3.4 million 个大于0.03 km2的湖泊，提供了1984~2019年全球的最大湖泊面积矢量文件。受限于不全面的洪泛平原定义，GLAKES没有完全去除错分的洪泛平原。此外，GLAKES无法较好地区分水稻田与湖泊（前者多分布于洪泛平原中，因季节性蓄水在occurrence底图中呈现Low occurrence）。本文利用SHIFT数据集(Zheng et al., 2024)和干旱指数数据集(Zomer et al., 2022)更新GLAKES的洪泛平原掩膜，去除GLAKES中的洪泛残留及错分的水稻田。SHIFT数据集是一个基于DEM映射的90 m分辨率的全球地貌洪泛区，比现有的洪泛区数据更全面。干旱区（Arid Index<0.2）湖泊较少受到洪水的影响 ，未对该区域的湖泊进行处理。和Pi等一致，本文使用30%的occurrence陆地掩膜处理SHIFT洪泛区的预测结果，获得了经掩膜处理的结果。Fig. 2显示了洪泛平原掩膜的良好效果，不仅有效地去除了错分的洪泛平原，使连片的湖泊群得以分离(Fig. 2a, b), 也有效去除了洪泛平原中聚集的水稻田(Fig. 2c, d)。

Fig. 2. Comparsion of GLAKES before and after applying the flood mask.

PLD是Wang等基于Circa-2015湖泊数据集及几个先进的水库数据集编制的全球湖泊掩膜，涵盖了全球约5.9 million个大于0.01 km2 的湖泊。PLD的主要来源为Circa-2015，提供了2015年前后的湖泊代表性面积。

2.2 Mapping global lakes using deep learning

U-Net是一种基于全卷积神经网络的语义分割模型，已被用于遥感的诸多领域。(Brandt et al., 2020)使用U-Net模型对西非撒哈拉沙漠和Sahel绘制了超过18亿的非森林冠（> 3平方米）地区，(Pi et al., 2022)基于(Brandt et al., 2020)开发的U-Net模型进行改进，绘制了全球3.4 million个湖泊。U-Net由编码器和解码器组成，能将深层与浅层特征进行融合，具有参数少、计算快、应用性强的特点。尽管不能输入任意大小影像，U-Net可通过重叠-切片 (overlap-tile) 策略对任意大的图像进行无缝分割。本文对Pi等的U-net模型进行改进，用于基于Sentinel-2数据的全球湖泊预测。

首先，我们在全球范围选择样本区域，制作深度学习训练所需的标签。为了更好地区分湖泊与其他地物，我们定义了5种代表性样本类别，类别具体为（1）普通的区域（Normal Regions, NR），通常情况下，湖泊形状偏圆形，在NDWI波段具有高值，在SWIR波段反射率低，在RGB波段中反射率较低，呈现深蓝色/黑色；（2）河边湖泊(Alongside Rivers, AR)，包含与河流形态相似的牛轭湖，需要仔细区分河流与湖泊；（3）建成区域（Built-up Regions, BR），部分建筑在NDWI波段与湖泊特征相似，需要加以区分；（4）冰川湖(Ice Lakes, IL)，冰川在NDWI波段与SWIR波段的特征与湖泊相似，需要加以区分；（5）盐湖（Salt Lakes,SL），盐湖具有特别的特征，其在RGB波段具有较高反射率；本文我们在全球范围内选择了799个代表性样本区域，并结合阈值提取方法与人工编辑方法获得样本区域对应的湖泊标签。样本的空间分布及大小分布如Fig. 3所示，其中，样本依据分层随机采样方法，被划分为训练（60％），验证（20％）和测试（20％）集。

Fig. 3. The spational distribution (a) and size range (b) of the sample regions, which divided into three subsets: training set (sky blue), validation set (orange), test set (green), the different markets represent 5 region types. (c) Total count and area of the sample regions in different region types.

随后，我们利用训练集和验证集对U-Net模型进行训练。样本的区域边界对于模型输入来说过大，因而本文使用了随机采样方法生成patchs（固定尺寸576×576），进而使用模型对patchs内的湖泊进行预测。针对Pi等的代码，本文主要做了如下修改：（1）使用随机采样策略生成训练样本时，调整不同大小区域的被选择概率，避免大样本区域欠采样，小的样本区域过采样的现象；（2）未采用Brandt等人的局部归一化方法，因为该方法会改变NDWI波段的特征；（3）引入,Recall, Precision及IoU协助模型评估。Recall代表标签中被正确提取的样本比例，precision代表预测结果中为正确样本的比例；IoU代表了…我们没有更改loss function及优化器的选择，但更改了具体的模型超参数设置（Supplementart Table 1）。训练过程中，我们同样依据模型在validation set上的loss error，保留最佳（loss error最小）的模型，并利用early stopping策略，提前终止模型训练（在本次实验中，是指50个epoch内 val loss没有降低）以避免模型过拟合。

获得最终模型后，即可对全球格网影像进行湖泊预测。由于格网影像过大，本文使用顺序裁剪方法获得patchs，将预测结果进行拼接。由于patch边缘上下文信息不足，预测准确性较低，本文舍弃了每一patch边缘100像素的预测值，提高湖泊预测的准确性。

获得初步预测结果后，随后进行了几步后处理操作。

(1) Ocean masking: 我们使用Openstreetmap 的ocean layer对初步预测结果进行ocean mask。 OSM是一个国际合作，旨在通过志愿者努力创建免费的地图数据源(Goodchild, 2007)，其shp格式的OSM海洋数据由<https://osmdata.openstreetmap.de/data/land-polygons.html> 网站获取；通过对所有中心点位于ocean layer内的海洋像素进行掩膜，能够去除大部分海洋残留。随后，我们将结果转换为矢量格式，进一步删除了所有与ocean layer相交的polygon。该步骤能够去除ocean layer未完全覆盖的海洋残留，同时也会删除与海洋相连的泻湖。由于泻湖的特性与内陆湖泊十分不同，本文数据集未考虑泻湖。

(2) River masking: 我们使用osm的river layer及OSMWL数据集对结果进行河流掩膜，并借助水库数据集（OSM的水库数据及GeoDAR数据集）避免对河流上的水库的错误掩膜（保留了绝大部分与水库相交的polygon）。OSM的全球数据可从<https://planet.openstreetmap.org/>获取，利用OSM标签可从中提取OSM的河流、水库数据。参考(Pi et al., 2022)的掩膜方法，本文计算出每一个polygon在初步河流掩膜（即删除所有中心点位于river layer中的像素）前后的面积比值，计算公式如下所示：

随后根据不同的情况进行处理，具体流程如Fig. 4所示。我们保留了与GLAKES相交且Area Ratio大于0.8的polygon(Fig. 4a)，从而保留了与河流相连的湖泊（如水库数据集未包含的水库）。不与GLAKES相交且 Area Ratio 小于0.8 的polygon基本上是河流残留(Fig. 4d)，被整个删除。其他情况下，mask后的结果被保留（Fig. 4b, c）。同时，我们进行了大量人工检查，以减少误差。

**Fig. 4.** Post-processing of river mask and the corresponding results.(a) Target polygons intersect with GLAKES with an area ratio ≥0.8; (b) Targets polygon not intersect with GLAKES with an area ratio ≥0.8; (c) Targets polygon intersect with GLAKES with an area ratio <0.8; (d) Targets polygon not intersect with GLAKES with an area ratio <0.8.

(3) Data-cutof: 所有小于0.005 km2的polygon被去除，因为该大小的polygon误差较大(Fig. 6)；

(4) Lake completion：本文从现有湖泊数据集中补充了部分漏分的湖泊。受限于感受野大小（10 m分辨率的516×516的patch覆盖至多33.17 km2陆地范围），本文模型会错误忽略部分大湖中类似河流的局部。而GLAKES利用30 m分辨率卫星数据进行湖泊提取，其大湖泊覆盖较全面。因此， 195084 个GLAKES湖泊（大于1km2的自然湖泊及所有水库）被用于大湖泊补全。同时，PLD数据集被用于补充干旱区湖泊。PLD的主要数据源Circa-2015数据集，已被验证在以大洋洲为代表的干旱地区有较好的预测效果。具体的，66627个干旱区中的PLD湖泊用于湖泊补全。湖泊补全操作前后的数量、面积变化如Table 1所示，181590个大湖泊被补全（断开的多个polygon得到连接），并新增了44736个湖泊。

Table 1 经补全的湖泊在数量、面积上的差异表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lake type | before lake completion | | after lake completion | |
| Count | Area | Count | Area after merge |
| The Salt Sea | 677 | 7825.966 | 1 | 41898.278 |
| New lake | - | 0 | 44,736 | 53,941.685 |
| Others | 440,523 | 2059548.841 | 181,589 | 2638976.622 |

2.3 Accuracy assessments

我们使用158个样本区域组成的测试集，对模型进行精度评定。我们进行了patchs尺度的分类别精度评价（Table 2）。总体上，模型表现较好的结果（Overall Accuracy 为 99.67%），但在部分区域上预测精度稍差。通过进一步进行目视检查（Fig. 5），发现模型在region AR上的 precision较低，是由于河流残渣的存在。我们计算了模型在test region中，河流掩膜前后的精度评价矩阵（Table 3），结果表明，河流掩膜有效提高了结果的precision。我们观察到模型在region SL中存在漏分现象（Fig. 5），recall仅72.71%（Table 2），lake completion 能够补全这部分盐湖。我们计算了lake-entity上，不同region及不同size group的精度评价结果（Fig. 6），随着湖泊尺度增加，标签的召回率及预测值的准确性在增加，小于0.005 km2的湖泊的平均召回率和平均准确性小于85%，因此本文选择0.005 km2作为数据集的最小湖泊尺寸，以减少误差。

Table 2 Accuracy assessments of our developed deep-learning algorithm at patch level with different region type.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Region type | Patch count | Overall Accuracy (%) | recall (%) | precision (%) | IoU (%) |
| NR | 798 | 99.75 | 93.00 | 96.26 | 89.71 |
| AR | 698 | 99.43 | 93.08 | 89.77 | 84.28 |
| BR | 308 | 99.87 | 95.78 | 90.83 | 87.29 |
| IL | 522 | 99.76 | 96.79 | 94.76 | 91.87 |
| SL | 98 | 99.53 | 72.71 | 95.60 | 69.80 |
| total | 2424 | 99.67 | 93.37 | 93.35 | 87.50 |

Fig. 5. The performance of the deep learning algorithm in test set. (a) Normal lakes; (b) Small river region; (c) Big river region; (d) Built-up region; (e) Ice lakes; (f) Salt lakes.

Table 3 The improvement of river mask operate in test regions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Overall Accuracy (%) | recall (%) | precision (%) | IoU (%) |
| Before river mask | 99.69 | 93.50 | 92.95 | 87.31 |
| After river mask | 99.73 | 93.40 | 94.72 | 88.78 |

Fig. 6. Validation in lake-entity of the deep-learning algorithm. (a) The Area of each label and its corresponding true predicted area. (b) The Area of each label and its corresponding true predicted area. (c) The mean recall of labels and mean precision of predicted polygons in different lake size groups.

3 Results

3.1 Global lake abundance and distribution

GLAKESplus囊括了~12 million个湖泊，总面积约3.4×106 km2。由于数据分辨率的提升及多波段的引入，GLAKESplus在小湖泊上的准确描绘有极大优势，更准确地反应了全球湖泊的分布。我们将湖泊分为small(<1 km2), medium (1~100 km2) and large (>100 km2)，可知全球湖泊面积由大湖泊占主导，贡献了56.1%的湖泊总面积，而小湖泊仅贡献了17.8%(Fig. 8a)。在数量上，全球湖泊则是由小湖泊主导，占据了98.3%(Fig. 7b)。Pi等指出小湖泊主导了过去四十年的湖泊变换，Williamson等指出小湖泊的温室气体排放巨大，说明了小湖泊的重要性。GLAKESplus提供了近4million个极小湖泊（<0.01 km2）的边界信息，贡献了全球湖泊数量的32.6%，减小了全球湖泊制图在小湖泊上的不确定性。在空间分布上，全球湖泊的49%或湖泊总面积的30%位于57°N以北的高纬度地区，主要分布在加拿大地盾、斯堪的纳维亚地区及西西伯利亚平原地区 (Fig. 7a)。这些地区在末次冰河时期冰川活动盛行，受更新世期间的冰川侵蚀作用形成了分布广泛的湖泊群。在57°N以南地区，湖泊数量逐渐减少，而直至36°N，湖泊面积没有显著减少，因为此处存在多个大湖（里海、五大湖、贝加尔湖、巴尔喀什湖，贡献了全球18%的湖泊总面积）(Fig. 7c)。在30°N、22°N附近，存在湖泊数量的小高峰(Fig. 7c)，主要由北美洲密西西比平原及长江中下游平原的小湖泊所贡献 (Fig. 7b)。热带地区湖泊数量及总面积都较少，但仍有几个密集分布区域。中国南方的沿海地区、柬埔寨及印度东部沿海地区，分布有大量的养殖鱼塘(Fig. 7b)。这些小湖泊对总湖泊面积的贡献不大，但分布于人口稠密地区，与人类的生产生活息息相关，十分的重要。南美洲亚马逊河流域及非洲的东非大裂谷有一些大湖泊，使得纬度剖面总面积在1~5°S之间存在一个小峰值。在经度剖面上，西半球的湖泊总面积、数量由加拿大地盾地区的密集湖泊主导；而东半球湖泊的总面积与数量的峰值不完全对应，湖泊面积由29~35°E的东非大裂谷湖泊及49~53°E的里海主导，湖泊数量则在66~120°E之间存在多个峰值(Fig. 7d)。

Fig. 7. Spational distribution of GLAKESplus. (a) Lake area density (total lake area/grid area). (b) Lake count per 1°×1° grid cell.

3.2 Comparison with other global lake datasets

我们将GLAKESplus与GLAKES、PLD进行比较，从而了解GLAKESplus在描绘全球湖泊分布上的优势与劣势。我们首先将三个数据集的湖泊按大小进行分组，统计湖泊总数及总面积，结果如Fig. 7所示。大于10000 km2的湖泊中，三个数据集数量一致(Fig. 7a)，但PLD的咸海因干涸被分为三个子湖泊，未被纳入该分组，导致对应湖泊总面积小于另外两个数据集(Fig. 7b)，又因密歇根-休伦湖的一分为二，湖泊数量没有减少。其余湖泊组中，GLAKESplus的湖泊总数量高于另外两个数据集，且差距随尺寸变小而增加，由于数据分辨率的提高，GLAKESplus 能绘制出更多的小湖泊(Fig. 8a)，在小湖泊绘制方面有显著优势。GLAKESplus的湖泊总面积则因湖泊补全操作(Fig. 8b)及新增的湖泊，皆高于另外两个数据集。在大于1 km2的湖泊中，GLAKES的湖泊数量及面积高于PLD。这是因为PLD主要数据来源为Circa-2015（提供2015年左右的湖泊平均水面积），而GLAKES提供近40年的最大湖泊边界，能绘制出更多的季节性湖泊及2015年时干涸的湖泊。在小湖泊（<1 km2）上，GLAKES湖泊总数量及面积皆最少，因其仅覆盖大于0.03 km2 的湖泊，且在小湖泊上有一定的漏分误差(Pi et al., 2022)。

Fig. 8. Total count and area of lakes across area size classes.

Fig. 9. Regional comparison among GLAKESplus , GLAKES and PLD. (a) Ponds in southeast China. (b) The Uyuni Salt Flat at the southern end of the South American Plateau. (c) Lakes by river. (d) Seasonal lakes in eastern Argentina. For (a-b), the left figures show the NDWI basemap and lake extend of GLAKES and PLD, the right figures show the RGB basemap and lake extend of GLAKESplus.

我们进一步分析了PLD与GLAKESplus在空间上的分布差异（Fig. 10）。在大于0.1 km2 的湖泊上，二者沿经度、纬度的面积分布没有较大差异，除PLD的湖泊总面积在60°E及45°N附近因咸海的干涸而偏低（Fig. 10a）。二者的湖泊总数量沿经纬度分布较为一致，GLAKESplus在120°E左右数量更多。在小于0.1 km2的湖泊上，GLAKESplus的总数量及总面积显著高于PLD (Fig. 10a, b, d, e)，不仅多绘制了小于0.01 km2的湖泊，也多绘制了0.01~0.1 km2的湖泊 。由于PLD没有考虑养殖鱼塘(Wang et al., 2023)，其在中国长江流域、印度东部及东南亚地区小湖泊显著少于GLAKESplus (Fig. 9c)。有少部分区域，PLD的湖泊数量、面积高于GLAKESplus。如部分流域地区 (Fig. 10c，Fig. 9c)，湖泊因季节性变化难以在多年平均影像上被模型捕捉到，并且部分牛轭湖因类似河流而被GLAKESplus忽略。此外，在干旱区边界未覆盖的一些地区，如阿根廷东部，湖泊因气候干旱而逐渐干涸，使得GLAKESplus的湖泊数量、面积比PLD低 (Fig. 10c,Fig. 9d)。但全球大部分地区，GLAKESplus都因其湖泊绘制的全面性而存在显著优势。

Fig. 10. 对比GLAKESplus与PLD的湖泊总面积及数量在空间分布上的差异（排除包含Garabogazköl lagoon的里海）。(a)每个经度上的湖泊总数量; (b)每个纬度上的湖泊总数量; (c)两个数据集的0.01~0.1 km2之间的湖泊在每个1度乘1度的格网内的数量、面积差异。(d) 每个经度上的湖泊总面积; (e) 每个纬度上的湖泊总面积; 在(a,b,d,e)子图中，大于0.1km2的湖泊与小于0.1 km2的湖泊被分别绘制在对向的两个坐标轴中，且坐标轴被经过拉伸，从而更好地显示。GLAKESplus由不同的颜色的堆叠柱状图表示，PLD由不同颜色及线宽的线型图表示。

4 Discussion

5 Conclusions

Supplementart Table 1 Keep hyperparameters tested and adjusted in the U-Net Model.

|  |  |
| --- | --- |
| Hyperparameters | Setting |
| Optimizer | Adadelta |
| Loss function | Dice loss |
| Batch size | 16 |
| steps\_per\_epoch | 691 |
| validation\_steps |  |
| Earlystopping patient | 50 |
| Epoch | 200 |
| Patch size | 576,576,6 |

**References**

Brandt, M., Tucker, C. J., Kariryaa, A., Rasmussen, K., Abel, C., Small, J., Chave, J., Rasmussen, L. V., Hiernaux, P., Diouf, A. A., Kergoat, L., Mertz, O., Igel, C., Gieseke, F., Schöning, J., Li, S., Melocik, K., Meyer, J., Sinno, S., … Fensholt, R. (2020). An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature*, *587*(7832), 78–82. https://doi.org/10.1038/s41586-020-2824-5

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

McFEETERS, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, *17*(7), 1425–1432. https://doi.org/10.1080/01431169608948714

Pi, X., Luo, Q., Feng, L., Xu, Y., Tang, J., Liang, X., Ma, E., Cheng, R., Fensholt, R., Brandt, M., Cai, X., Gibson, L., Liu, J., Zheng, C., Li, W., & Bryan, B. A. (2022). Mapping global lake dynamics reveals the emerging roles of small lakes. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-33239-3

Wang, J., Pottier, C., Cazals, C., Battude, M., Sheng, Y., Song, C., Sikder, M. S., Yang, X., Ke, L., Gosset, M., Oliveira, R. R. A., Grippa, M., Girard, F., Allen, G. H., Biancamaria, S., Smith, L., Crétaux, J.-F., & Pavelsky, T. M. (2023). *The Surface Water and Ocean Topography Mission (SWOT) Prior Lake Database (PLD): Lake mask and operational auxiliaries* [Preprint]. Preprints. https://doi.org/10.22541/au.170258987.72387777/v1

Zheng, K., Lin, P., & Yin, Z. (2024). SHIFT: A spatial-heterogeneity improvement in DEM-based mapping of global geomorphic floodplains. *Earth System Science Data*, *16*(8), 3873–3891. https://doi.org/10.5194/essd-16-3873-2024

Zomer, R. J., Xu, J., & Trabucco, A. (2022). Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01493-1