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

Beihui Hu1, and Lian Feng1

1 Introduction

湖泊和水库（以下简称 “湖泊”）在全球水文和生物地球化学水循环及碳循环中发挥着重要作用（Lehner & Döll, 2004; Verpoorter et al, 2016）。它们的作用与其几何特征密切相关(Messager et al., 2016).。由于其对气候变化和人类活动的敏感性，全球湖泊不断发生变化(Pi et al., 2022; Williamson et al., 2009)。掌握全球湖泊的空间分布和变化对相关的地球系统过程研究及水资源调控至关重要。

因为湖泊分布广泛，形态各异，数量庞大，全球湖泊制图存在着挑战。现有数据集中的GLWD、HydroLAKES，是由多个数据源编制而成的。这些数据集不能很好地体现全球湖泊分布，因其数据源没有统一的分辨率及制图时间。遥感因其覆盖面广、时效性强、信息量大的优势，已成为大范围湖泊观测的主要方法。粗分辨率卫星传感器如AVHRR卫星的AVHRR(空间分辨率1001m)及MODIS的Terra/Aqua（空间分辨率250m至1km仅适用于大湖泊的研究。中等分辨率卫星Landsat自2008年免费对公众开放后，被广泛应用于全球湖泊制图研究。如Verpoorter等处理了2000年左右Landsat7上的 ETM+传感器的图像制作了数据集GLOWABO，绘制了全球大于0.002 km2的湖泊。Sheng等利用2015年左右的Landsat8 OLI传感器 30m数据，计算NDWI水体指数，制作了Circa-2015数据集，提供大于0.004 km2的全球湖泊代表性水面积。因为地表水体形态多样，GLOWABO及Circa-2015可能存在错分及漏分误差，并且他们的制图数据仅为一个特定时期。

GSW产品是基于Landsat数据的地表水产品。Pekel等利用专家系统对每幅Landsat影像进行水分类，其获得的水概率密度产品很好地表现了地表水的分布。Pi等利用深度学习方法处理GSW水概率产品，对全球湖泊在1984至2019年的面积变化进行研究，揭示了了小湖泊的重要作用。Pi等的湖泊数据集GLAKES涵盖全球3.4 million个大于0.03 km2的湖泊，提供研究时段内的最大湖泊边界及分时段的加权面积。

尽管现有的基于Landsat的全球湖泊数据集的最小单元为0.002 km2，小湖泊的空间分布仍未得到准确掌握。尽管小湖泊对全球湖泊总面积的贡献较小，但其占湖泊总数量的主导地位。已有研究表明小湖泊是重要的内陆水域碳通量来源，其面积变化更加剧烈。因此，有必要制作一个覆盖更多小湖泊的全球湖泊数据集。而随着高分辨率卫星的投入使用，这一目标成为可能。相比较于昂贵的商业卫星（如SPOT卫星、QuickBird卫星），免费提供低至10米空间分辨率数据的Sentinel-2卫星在大范围湖泊监测中有更大的优势。目前，已有基于Sentinel-2的区域性湖泊观测研究，如…但全球湖泊10米分辨率湖泊制图所需的海量遥感数据对计算机存储及处理方法都有较高要求，在这方面还没有相关的研究。

近年来出现的遥感云计算平台使海量遥感数据处理成为可能。其中，常见的Google Earth Engine平台提供了自2017年3月28日以来的Sentinel-2 L2A产品。利用GEE平台的预处理方法，能够高效处理全球卫星影像，并减少对本地存储的要求。此外，新兴的深度学习技术具有准确率高、速度快、自动化程度高的优点，在大范围遥感观测研究中有巨大优势。如Brandt等使用亚米级分辨率卫星影像和深度学习技术绘制了西部撒哈拉地区超过18亿株冠幅尺寸大于3m2的树。Pi等利用深度学习技术绘制了全球三百多万个大于0.3 km2的湖泊。深度学习模型不仅能学习地物的光谱特征，还能学习其几何特征。相较于传统的湖泊识别方法，其在区分湖泊与其他地表水体方面具有极大优势。

基于先前的研究，本研究的目的是利用Sentinel-2遥感数据及深度学习方法，制作一个10 m分辨率的全球湖泊数据集。具体的研究目的包括：1）训练一个适用于Sentinel-2全球湖泊制图的语义分割模型；2）对全球湖泊进行绘制，提供一个比现有数据集覆盖更多小湖泊，边界更精准的全球湖泊数据集；3）对全球湖泊分布进行分析，并将结果与其他数据集进行对比。

2 Materials and methods

GLAKESplus的制作过程如下：1) 影像处理，研究时段内的Sentinel-2数据的NDWI、红、绿、蓝及近红外波段的逐像素平均结果被计算，随后被分片下载至本地存储；2) 样本准备，全球范围内的样本区域及湖泊标签被创建，用于制作训练数据集（由训练集、验证集、测试集组成）；3)模型应用，U-Net模型被训练来从预处理影像中提取湖泊特征，并预测出了原始全球湖泊分类图；4)后处理，我们进行了若干后处理步骤来减少原始全球湖泊分类图中错分、漏分的现象，最终生成了GLAKESplus湖泊多边形。技术路线如Fig. 1所示。

Fig. . Flowchart for developing the GLAKESplus dataset.

2.1 Sentinel-2 and preprocessing

Sentinel-2是欧洲航天局( ESA )哥白尼计划的下的一个地球观测任务，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。Sentinel-2的B2(蓝光)、B3(绿光)、B4(红光)及B8 (近红外)波段空间分辨率为10米，可以展示精细的湖泊边界。其近红外波段（B11,B14）空间分辨率为20 m,其余波段空间分辨率为60 m。NDWI是一个常用的水体增强指数，其通过对绿光、红光波段的组合运算，抑制植被信息，增强水体信息。在Sentinel-2中的计算公式为。我们选择了NDWI波段及三个可见光波段（B2,3,4），一个短波红外波段B11波段（重采样至10 m）用于后续的湖泊制图。建成区的建筑与水体在NDWI波段易被混淆，但在短波红外波段有着显著差异。因此，纳入短波红外波段有利于对建成区湖泊的识别。我们在GEE上获得了2017年3月28日至2022年4月10日的所有云像素比例低于60%的Sentinel-2 L2A影像，并进行了若干预处理操作。

首先，哨兵云概率产品(S2 Cloud Probability, S2Cloudless)被用去云像素。S2Cloudless是使用 sentinel2-cloud-detector 库创建的，其提供了每幅影像的单个像素的云概率。根据经验，对所有云概率高于50%的像素进行掩膜操作，并将近红外波段的暗像元与云投影取交集，去除云阴影像素。并且，部分冰雪像素被通过场景分类图层 (Scene Classification Layer, SCL)去除。场景分类图层是L2A产品的一个波段，提供对云、云阴影、冰雪、水体的初步的分类。随后，我们计算了每幅影像的NDWI指数并将所有影像进行逐像素平均，获得一个平均影像。该操作能获得湖泊的平均边界，减少湖泊季节性变化的影响，同时消除云残留、藻华、泥沙等因素的干扰。最终，经过处理的近10TB 的Sentinl-2平均影像被分片下载至本地计算机，作为后续湖泊识别的源数据。

2.2 Mapping global lakes using deep learning

深度学习指通过**多层神经网络**来自动学习和提取输入数据中的复杂特征，已经成为过去几十年人工智能领域的主要驱动及过程(Brandt et al., 2020)。深度学习领域出现了一系列功能强大的模型，其中 U-Net 模型及其变体在语义分割领域得到了广泛应用。U-Net是一种基于全卷积神经网络的语义分割模型，由主干特征提取网络、加强特征网络与预测网络构成。网络左侧可视为一个编码器 (Encoder)，右侧可视为一个解码器 (Decoder)。通过编码与解码，U-Net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。结合重叠-切片 (overlap-tile) 策略，U-Net可对任意大的图像进行无缝分割，已被广泛应用于遥感领域。如Brandt等调整了U-Net模型，将其应用于基于高分辨率影像的树木识别。随后，Pi对Brandt等的模型进行调整，根据空间分辨率为 30 米的 GSW 数据，获得了良好效果。因此，我们在Pi等的代码上进行细微调整，将U-Net应用于基于Sentinel-2影像的全球湖泊10 m分辨率制图。

2.2.1 Sample preparation

我们在全球范围选择具有代表性的样本区域，用于制作训练U-Net模型所需的标签。针对每一个样本区域，我们通过基于NDWI波段的阈值分割法获得初步标签，阈值由人工确定。随后，非常耗时的人工修改被应用，以去除初步标签中的错分、漏分误差，获得最终标签。在Sentinel-2影像中，湖泊的NDWI值通常比背景高，在RGB、SWIR波段反射率较低，形状偏圆形，表面平坦，易于和背景区分。易于区分的区域被我们标记成（1）普通区域（Normal Regions, NR）；此外，我们还观察到几种需要仔细辨认的区域，分别为（2）河边湖泊 (Alongside Rivers, AR)，分布有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），建筑及其阴影在NDWI中易被误判为湖泊，需要结合其他波段进行剔除；（4）冰川湖(Ice Lakes, IL)，冰川在NDWI波段及SWIR波段的特征与湖泊相似，需要仔细区分冰川湖的边界；（5）盐湖（Salt Lakes,SL），盐湖通常分布于内陆干旱和办干旱地区，NDWI分割的阈值偏低，在其它几个波段反射率较高；最终，我们选择并绘制了799个的样本标签，并按照分层随机采样方法，将其划分为训练（60％），验证（20％）和测试（20％）集。样本区域的空间、大小分布如Fig. 3所示，应当指出的是，样本类型指的是这个区域中的主要水文特征。

Fig. 2. 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.

2.2.2 Model application

训练深度学习网络的过程，就是最小化损失函数，使模型预测结果与真实标签尽可能接近的过程。我们采用了与Pi等一致的损失函数和优化算法。训练过程中，模型在训练集上的梯度被计算，从而依据优化算法更新模型参数。验证集则被用于保存最佳模型（具体的，我们保留在验证集上loss error最小的模型）。由于GPU内存的限制，模型输入被设定为576×576×5。由于样本区域对模型输入来说过大，因此一个随机采样方法被用于从训练集、测试集中生成与模型输入大小一致的patches。每个样本区域被随机选中的概率与其大小成正比，以避免对大样本的欠采样和对小样本的过采样；此外，early stopping策略被用来避免模型的过拟合，当连续50个epoch的val loss没有降低时，训练被提前终止。训练的具体的超参数设置如Supplementart Table 1所示。

获得最终模型后，我们对全球湖泊进行预测，获得了初步预测结果。由于全球格网影像过大，其被从左到右，从上到下移动的sliding window裁剪为与模型输入一致的patchs，输入模型进行预测。每个patch的输出结果被拼接成整幅影像的结果。由于单个patch边缘上下文信息不足，预测准确性较低，其边缘100像素宽度的预测值被舍弃，进而提高湖泊边界的准确性。

2.2.3 Post processing

结合多源辅助数据集，几步后处理操作被应用。受限于相对小的输入大小，深度学习模型难以区分局部特征与大湖泊类似的海洋。因此，我们使用OSM数据集对初步预测结果进行掩膜，以去除海洋残留。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Coastline data sets（包括代表全球land area的land polygon及代表海洋的water polygon）可以从 <https://osmdata.openstreetmap.de/data/>获得。 首先，所有不在Land polygon海洋像素被去除。随后，所有与ocean polygon相交的polygon被进一步删除，以去除OSM未完全覆盖的海洋残留。同时，与海洋相连的泻湖也被删除。泻湖没有被纳入我们的考虑范围，因其特性与内陆湖泊十分不同。尽管深度学习模型能很好地区分湖泊与小河流，其在区分湖泊与大河宽的河流方面同样由于相对小的输入大小而存在局限性。大河宽的河流在全球河流的数量占比很低，且已有诸多大河宽的河流数据集，如Global River Widths from Landsat (GRWL)数据集。结合GRWL数据集及OSM的河流图层，能够有效去除结果中的河流残留。OSM的河流、水库数据由OSM的全球数据（<https://planet.openstreetmap.org/>）中提取。本文参考(Pi et al., 2022)的方法进行河流掩膜。首先，我们利用水库数据集（OSM的水库数据及GeoDAR数据集）保留了河流上的水库。此外，所有中心点位于river layer中的像素被初步去除。随后，每一个polygon在初步河流掩膜前后的面积比值被计算，计算公式如下所示：

当越接近1，表明这个polygon越可能是与河流相连的湖泊，当AR接近0，表明其更可能是河流残留。根据目视检查，我们保留了与GLAKES相交且Area Ratio大于0.8的polygon (Fig. 3a)，从而保留与河流相连的湖泊（包括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被整个删除，因其基本上是河流残留(Fig. 3d)。其他情况下，初步掩膜后的结果被保留（Fig. 3b, c）。我们对各大流域进行了大量人工检查，以减少误差。

Fig. 3.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.

受限于底图的分辨率，模型的预测精度随湖泊的减小而降低。基于测试集的精度评价结果，我们设定最小湖泊面积阈值 ，删除了所有小于阈值的polygon；

经过以上操作，初步预测结果中的错分误差被基本消除。受限于输入大小，深度学习模型难以分辨大湖泊类似河流的局部特征，从而导致部分大湖泊存在漏分现象。此外，深度学习模型难以从平均影像中全面捕捉干旱区的湖泊。为了获得更全面的湖泊数据集，我们进一步进行了湖泊补全操作，利用GLAKES和Circa-2015湖泊数据集补全缺漏的湖泊。GLAKES提供了1984~2019年全球的最大湖泊面积矢量文件，适用于大湖泊补全。Circa-2015提供了2015年左右的全球代表性水面积已被证明对大洋洲高度动态的湖泊的良好捕捉 (Sheng et al., 2016)，可以补充干旱区湖泊，其数据从PLD数据集中获得。我们利用干旱指数数据集 (Zomer et al., 2022)定义了干旱区范围（Arid Index<0.2）。

首先，我们对GLAKES数据集进行了处理，以去除其中因不全面的洪泛平原定义而包含的洪泛平原。我们利用SHIFT数据集 (Zheng et al., 2024) 对GLAKES非干旱区的湖泊进行洪泛平原掩膜。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。本文使用和Pi等一致的30% occurrence阈值去除GLAKES湖泊包含的洪泛平原，获得良好效果。更新的洪泛平原掩膜操作不仅有效地去除了GLAKES中错分的洪泛平原，使连片的湖泊群得以分离 (Fig. 4a, b), 也有效去除了常分布于洪泛平原中的因季节性蓄水而被错分为湖泊的水稻田 (Fig. 4c, d)。

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

最终，经洪水掩膜的GLAKES数据集中大于1 km2的自然湖泊及所有水库，及66627个位于干旱区中的面积大于0.01 km2的Circa-2015湖泊，被合并至初步预测结果中。经过大量人工检查和修改，得到了最终预测结果GLAKESplus数据集。

3 Results

3.1 Accuracy assessments

158个样本区域组成的独立测试集，被用于对模型的评估。我们使用的评价指标包含Recall, Precision及IoU。Recall是正确湖泊预测占真实湖泊标签的比例，precision是正确湖泊预测占所有湖泊预测的比例；F1 Score则是Recall和Precision的harmonic average，接近这两个数字的最差值；IoU代表正确样本占标签与预测的并集的比例。因为这是二分类问题，我们没有使用mIoU,并且，我们没有使用overall accurracy，以避免大量正确的负样本（即被正确分类为非湖泊）影响精度评价结果。首先，我们利用顺序采样方法从测试集中获取2424个576×576大小的patches，进行了patches尺度的分类别精度评价（Table 2）。总体上，模型表现出较好的结果，IoU达到87.5，其他指标高于92.7。在不同区域类别上，模型表现稍有差异。在region AR上，由于河流残渣的存在（Supplementary Figure 1c）,模型的 precision稍低，为89.77（Table 2）。通过河流掩膜操作有效去除河流残留，模型在测试集上的表现得到提高（Table 3）。在region SL上，模型的recall仅72.71%（Table 2），对盐湖有一定的漏分（Supplementary Figure 1f）。盐湖占全球湖泊的数量较少，多分布于干旱区，能够通过lake completion 操作补全。

Table 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 |

Table 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 |

在lake-entity上，正确湖泊分类与标签及所有湖泊预测皆表现出很高的相关性，R方都是0.98，斜率大于0.92（Fig. 5a）。被河流掩膜操作去除的预测值图标为灰色，其真实湖泊预测值趋近0，表明了河流掩膜操作的有效性。随着尺度变小，正确湖泊分类与标签的相关度降低（Fig. 5a），且不同尺寸组中标签的平均召回率及预测值的平均准确性在减小（Fig. 5b）。因此，我们选择0.005 km2作为数据集的最小湖泊尺寸，有效减少了误差。

Fig. . 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.

3.2 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. 6所示。大于10000 km2的湖泊中，三个数据集数量一致 (Fig. 6a)，而PLD的总面积稍小于GLAKESplus及GLAKES(Fig. 7b)，主要由咸海的差异导致。其余湖泊组中，GLAKESplus的湖泊总数量皆高于另外两个数据集，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. 7a)，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. 7b)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 . Comparisons among GLAKESplus, GLAKES and PLD in logarithmic space.

Fig. . 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. 9）。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. 9a）。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. 9a, 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. 8c)。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. 9c，Fig. 8c)，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. 9c,Fig. 8d)。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。

3.3 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. 6a)。在数量上，全球湖泊则是由小湖泊主导，占据了98.3%(Fig. 6b)。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4 million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. 6a)。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. 6c)。在30°N、22°N附近，存在湖泊数量的小高峰(Fig. 6c)，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. 6b)。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. 6b)。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. 6d)。

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

4 Discussion

5 Conclusions

相较现有研究，GLAKESplus不仅提供了更精细的湖泊几何边界，也对小湖泊的捕捉更加全面，有助于内陆湖泊碳通量的准确评估，能为气候变化、内陆水域的生物地球化学研究提供数据支持，对全球水资源保护有重要意义

Supplementary Figure . 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.

Supplementary Figure 2. 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。

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

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13603

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

Sheng, Y., Song, C., Wang, J., Lyons, E. A., Knox, B. R., Cox, J. S., & Gao, F. (2016). Representative lake water extent mapping at continental scales using multi-temporal Landsat-8 imagery. *Remote Sensing of Environment*, *185*, 129–141. https://doi.org/10.1016/j.rse.2015.12.041

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

Williamson, C. E., Saros, J. E., Vincent, W. F., & Smol, J. P. (2009). Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnology and Oceanography*, *54*(6part2), 2273–2282. https://doi.org/10.4319/lo.2009.54.6\_part\_2.2273

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