remote sensing

MDPI

Article

Map of Land Cover Agreement: Ensambling Existing Datasets for Large-Scale Training Data Provision

Gorica Bratic *, Daniele Oxoli and Maria Antonia Brovelli

Department of Civil and Environmental Engineering, Politecnico di Milano, 20133 Milan, Italy; daniele.oxoli@polimi.it (D.O.); maria.brovelli@polimi.it (M.A.B.)

* Correspondence: gorica.bratic@polimi.it

Abstract: Land cover information plays a critical role in supporting sustainable development and informed decision-making. Recent advancements in satellite data accessibility, computing power, and satellite technologies have boosted large-extent high-resolution land cover mapping. However, retrieving a sufficient amount of reliable training data for the production of such land cover maps is typically a demanding task, especially using modern deep learning classification techniques that require larger training sample sizes compared to traditional machine learning methods. In view of the above, this study developed a new benchmark dataset called the Map of Land Cover Agreement (MOLCA). MOLCA was created by integrating multiple existing high-resolution land cover datasets through a consensus-based approach. Covering Sub-Saharan Africa, the Amazon, and Siberia, this dataset encompasses approximately 117 billion 10m pixels across three macro-regions. The MOLCA legend aligns with most of the global high-resolution datasets and consists of nine distinct land cover classes. Noteworthy advantages of MOLCA include a higher number of pixels as well as coverage for typically underrepresented regions in terms of training data availability. With an estimated overall accuracy of 96%, MOLCA holds great potential as a valuable resource for the production of future high-resolution land cover maps.

Keywords: training data; high-resolution land cover; global land cover; machine learning; deep learning; satellite image classification; classification accuracy assessment

Citation: Bratic, G.; Oxoli, D.; Brovelli, M.A. Map of Land Cover Agreement: Ensambling Existing Datasets for Large-Scale Training Data Provision. *Remote Sens.* **2023**, *15*, 3774. https://doi.org/10.3390/rs15153774

Academic Editors: Stefano Nativi, Gregory Giuliani, Joan Masó and Paolo Mazzetti

Received: 28 June 2023
Revised: 19 July 2023
Accepted: 28 July 2023
Published: 29 July 2023

1. Introduction

In today's world, precise and comprehensive land cover (LC) mapping is becoming increasingly crucial for sustainable development and well-informed decision-making. Beyond its relevance in climate studies [1], LC information finds utility in other fields as well. For instance, in ecology, LC data aids in estimating habitat fragmentation and predicting International Union for Conservation of Nature (IUCN) Red List categories for species [2]. Additionally, LC serves as a crucial variable in hydrological investigations, as exemplified by studies conducted in the upper Crepori river basin in Brazil and the Gumara catchment in Ethiopia [3,4].

The applications of LC data extend to monitoring various phenomena across different regions. Examples include monitoring the desertification process in the Qubqi desert in China [5], tracking urbanization progress in Abuja, Nigeria [6], and observing agricultural expansion in the Mato Grosso state of Brazil [7]. These cases illustrate the diverse range of uses for LC data in monitoring and understanding our changing environment.

The inclusion of open data policies by some providers of satellite imagery has undeniably accelerated the progress of LC mapping [8–10]. This favorable development, along with advancements in computing capabilities and satellite technologies, has made significant contributions to the field. Nevertheless, there are still persistent challenges in the domain of LC mapping.

To fully harness the potential of satellite Earth observation resources for land cover mapping, it is crucial to address a significant challenge: the availability of appropriate

training data. Specifically, the effectiveness of machine learning (ML) algorithms used to generate land cover maps relies heavily on the quality and relevance of the training data [11,12].

In the case of extensive classification tasks such as global high-resolution land cover (HRLC) mapping, the requirements for training datasets become even more demanding. This is because the training data needs to encompass vast geographical areas and offer representative samples with a high level of detail that can capture the diverse landscape characteristics worldwide. Furthermore, deep learning techniques, which are current state-of-art-techniques for LC classification, typically require larger training datasets compared to classical ML techniques [13–15].

Dimitrovski et al. [16] summarized 22 open-access training datasets used for deep learning approaches. The datasets comprise image chips of different dimensions primarily obtained from aerial imagery, supplemented by a limited number sourced from satellite imagery. The biggest dataset—among the ones revised—is Big Earth Net which has samples covering approximately 750,000 km^2 which are located only in Europe [17]. There are also datasets with global coverage such as Resisc45 [18] and MLRSNet [19] but covering smaller areas than Big Earth Net—470,000 km^2 and 182,000 km^2, respectively.

The practice of global HRLC producers to obtain training data includes photo-interpretation, utilization of existing LC data at various resolutions, and sometimes a combination of the two [20]. DynamicWorld project, the first project for near-real-time global LC mapping, generated its own training dataset of 5 billion 10 m pixels and released it publicly [21]. The dataset was derived by the photo-interpretation of Sentinel-2 images, dominantly performed by non-expert annotators. The same training was reused by Esri LC [22]. The collection of training data was based on a photo interpretation for datasets such as Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) [23–25], World Settlement Footprint (WSF) [26,27], Global Surface Water (GSW) [28], and Forest Non-Forest (FNF) [29]. Various HRLC production projects utilized existing LC data in different ways to support their training dataset collection. For instance, the GlobeLand30 (GL30) dataset allowed photo-interpreters to refer to existing LC datasets during their work [30]. In the case of the initial version of Global Human Settlements Built-up (GHS BU) datasets, a combination of low-resolution LC (LRLC) and HRLC datasets were employed, with a weighted voting system favouring the HRLC data [31]. HRLC and medium-resolution LC (MRLC) data, along with photo-interpretation, were utilized to derive the Tree Canopy Cover Dataset [32]. The Global Mangrove Watch (GMW) dataset combined both HRLC and LRLC datasets [33,34]. The European Space Agency's (ESA) World Cover dataset used existing MRLC and HRLC data to extract training data, although the specific method employed remains unclear [35]. Initially, the Global Cropland dataset relied on photo-interpreted samples [36]. The generated LC data was then sampled to obtain training data for subsequent iterations until satisfactory accuracy was achieved. The Global Land Cover with a Fine Classification System at a 30-m resolution (GLC_FCS30) dataset utilized refined MRLC data, obtained through a specific procedure that considers only homogeneous samples [37].

It is apparent that HRLC producers were aiming to incorporate existing LC data into their training data extraction process, likely due to the high cost associated with global data collection. However, they did not always consider the reliability of training samples derived from existing HRLCs, as observed in the case of GMW.

Among the listed global existing HRLCs, the highest overall accuracy (OA), equal to 86%, was achieved for GL30 and the first release of Esri LC [38,39]. The details of GL30 accuracy are not published, while the second release of Esri LC merged Grass and Scrub classes into a single class—Rangeland to compensate for the low accuracy of these classes in the first release. Although achieving an accuracy of 86% is a noteworthy advancement for HRLC products, it is evident that there is room for further enhancements, especially in specific classes [40].

In this paper, we present the training benchmark dataset that was generated by borrowing two concepts of training sample generation techniques: reuse of existing data [31–33,35–37] and consensus among multiple annotators in the case of photo interpretation [23,41]. During the human labeling of training samples, human error is often mitigated by having multiple annotators. If there is no consensus among them, or at least among the majority, the sample is rejected.

Adhering to the above principles, we reuse existing HRLC datasets, but only those portions in which there is exclusive consensus among multiple datasets. From a practical standpoint, multiple HRLCs are combined using the intersection method to retain only the areas where all datasets agree on LC classes while disregarding areas of disagreement. Accordingly, the dataset obtained is named Map Of Land Cover Agreement (MOLCA). The main purpose of MOLCA is to serve as a reference training dataset, from which to extract samples that are functional for the creation of new HRLC maps. MOLCA was designed to provide training samples mainly for large-scale HRLC mapping using ML and deep learning techniques, which typically demand extensive training data for satellite imagery classification. This dataset was produced within the Climate Change Initiative HRLC (CCI HRLC) project of the ESA. MOLCA has 117 billion 10 m pixels (11.7 million km^2) distributed over an area of 19 million km^2. Classes included in the MOLCA legend are Bareland, Built-up, Cropland, Forest, Grassland, Shrubland, Water, Wetland, and Permanent ice and snow, which depicts LC in the period between 2016–2020. The accuracy estimate of MOLCA shows an OA of 96%. MOLCA offers distinct advantages over alternative methods of training data collection, including a substantially larger number of available pixels and coverage for regions that are frequently underrepresented in existing benchmark training datasets, such as Africa and Siberia [17,42–49]. Standing on the analyzed literature, MOLCA outperforms other existing open-access training datasets in terms of spatial coverage and precautions taken to ensure a high level of accuracy, due to the consideration of multiple—instead of individual—HRLC maps for the generation of training samples. These key features of MOLCA are promising to foster its use in future HRLC map production. Furthermore, the availability of MOLCA as open data further enhances its potential for widespread use.

The structure of this paper is as follows: Section 2 outlines the region considered for MOLCA generation, input datasets, data generation concepts and methodology, and the validation approach. Section 3 presents statistical information and the accuracy evaluation of the generated dataset. The analysis and interpretation of the results are discussed in Section 4, while the concluding remarks are provided in Section 5.

2. Materials and Methods

MOLCA was produced in the context of the CCI HRLC project of the ESA. The region of interest for the project encompasses three macro-regions of the world: Amazon, Siberia, and Sub-Saharan Africa (see Figure 1).

The region of interest extends over 19,163,868 km^2–4,526,839 km^2 in the Siberia, 6,203,824 km^2 Amazon, and 8,433,205 km^2 Sub-Saharan macro-region. The objective of the CCI HRLC project was to determine the impact of the increased spatial resolution of land cover data on climate models. Besides the selected regions being only partially represented by existing LC training datasets, they are also landmarks for climate change. For these reasons, they were selected for the first implementation of MOLCA.

2.1. Input Datasets

In the derivation of MOLCA, multiple global HRLCs were used as common input across all three regions of interest. However, within each region, an additional regional HRLC was incorporated into the MOLCA computation. The global datasets employed included two general-purpose HRLCs, namely FROM-GLC and GL30, along with two thematic HRLCs specific to the built-up class (WSF and GHS BU Sentinel-1—GHS BU S1NODSM), one thematic HRLC for water (GSW), and one thematic HRLC for forests

(FNF), as indicated in Table 1. As for the regional HRLCs, MapBiomas was used for the Amazon region, CCI Africa Prototype for Africa, and ESA DUE (Data User Element) GlobPermafrost for Siberia. All regional datasets fall under the general type.

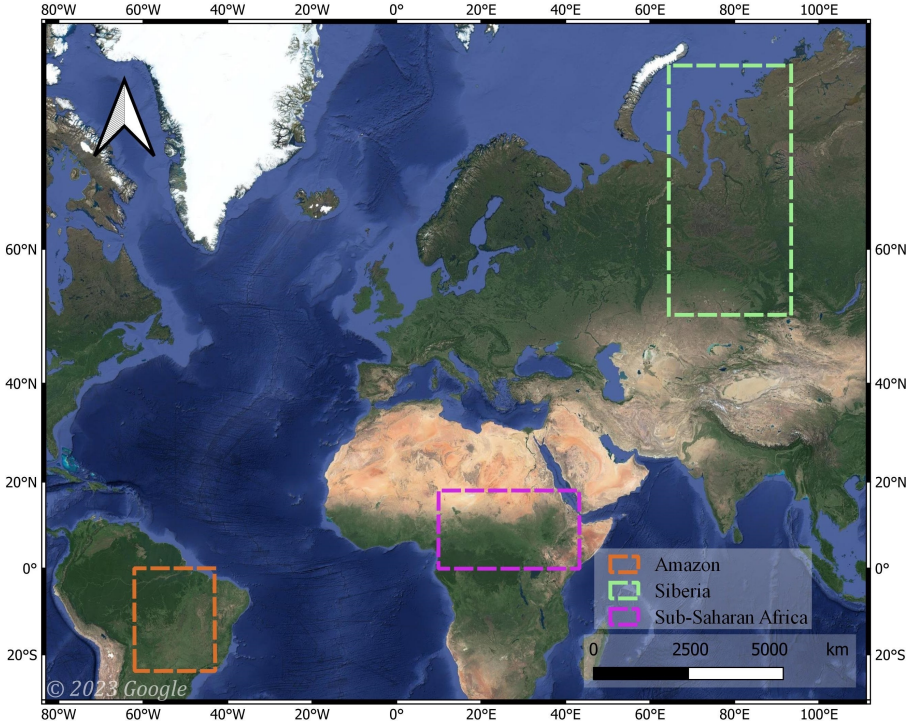

Figure 1. Regions where MOLCA data are produced.

The baseline year for these datasets ranged from 2016 to 2020, and the spatial resolution varies from 10 m to 30 m. The most used CRS is WGS84, while a few datasets are supplied in UTM or Web Mercator (see Table 1).

Table 1. Existing global and regional HRLCs used for MOLCA creation in the three regions of interest.

Existing HRLC	Baseline Year	Coverage	Resolution	CRS	Type
FROM-GLC	2017	Africa, Amazon, Siberia	10 m	WGS84	General
GL30	2020	Africa, Amazon, Siberia	30 m	UTM	General
GHS BU S1NODSM	2016	Africa, Amazon, Siberia	20 m	Web Mercator	Thematic built-up
WSF	2019	Africa, Amazon, Siberia	10 m	WGS84	Thematic built-up
FNF	2018	Africa, Amazon, Siberia	25 m	WGS84	Thematic forest
GSW	2019	Africa, Amazon, Siberia	30 m	WGS84	Thematic water
MapBiomas	2019	Amazon	30 m	WGS84	General
CCI Africa Prototype	2016	Africa	20 m	WGS84	General
ESA DUE GlobPermafrost	2016	Siberia	20 m	UTM	General

In this work, we utilized the 2017 map from FROM-GLC, which is a collection of irregular time series of general-purpose land cover (LC) maps developed by Tsinghua University [23–25]. The map has a resolution of 10 m and consists of 10 classes in its legend. It is provided in World Geodetic System 1984 (WGS84) Coordinate Reference System (CRS) in the form of 10° × 10° tiles (http://data.ess.tsinghua.edu.cn, accessed on 28 June 2023). The reported OA of this map is 73%.

As for the GL30 dataset, it is a regular time series of general-purpose LC maps at a resolution of 30 m, developed by the National Geomatics Center of China (NGCC) [30].

The legend of GL30 consists of 10 classes. For this work, we used the 2020 product version. The reported OA for this map is 86%, as mentioned on the GL30 website (http://globeland30.org, accessed on 28 June 2023). The distribution of the GL30 product is based on the Universal Transverse Mercator (UTM) projection. The tile size of GL30 varies depending on the location, with most tiles (between 60°N and 60°S) being 5° × 6° in size, although some tiles can be 5° × 12° or even larger.

A comprehensive set of thematic maps called the Global Human Settlement Built-up (GHS BU) distinguishes between built-up and non-built-up surfaces [50,51]. These maps were developed by the Joint Research Centre (JRC) of the European Commission. Various GHS BU products exist, each one differing in terms of input imagery, baseline year, and production method. In this study, the GHS BU S1NODSM product, which is based on Sentinel-1 imagery from 2016, was employed. It consists of two classes: Built-up and non-built-up. The product is distributed as a compressed file folder containing 2° × 2° tiles that cover the entire globe (https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL, accessed on 28 June 2023). The original CRS of the tiles is Web Mercator projection (EPSG:3857). The accuracy of GHS BU S1NODSM is described qualitatively in comparison to another LC dataset [50].

Another thematic LC product specifically focused on built-up areas is the WSF from the German Aerospace Center—DLR [27]. It encompasses two classes: Settlements and non-settlements. The product includes two maps with a spatial resolution of 10 m, representing 2015 and 2019. The 2019 map was utilized in this research. WSF is available as 2° × 2° tiles in the WGS84 CRS (https://download.geoservice.dlr.de/WSF2019, accessed on 28 June 2023). The WSF map for 2019 exhibits an OA of 84% and a Kappa value of 0.65, although information regarding User's Accuracy (UA) and Producer's Accuracy (PA) is currently unavailable [52].

The GSW family comprises a collection of multi-temporal thematic LC maps that focus on inland water bodies [28]. Produced by JRC, these annual maps span 37 years, from 1984 to 2021. The GSW product offerings include various aspects such as monthly water history, seasonality, yearly history, water occurrence, change intensity, recurrence, transitions, maximum water extent, monthly recurrence, and metadata. They are available for download at https://global-surface-water.appspot.com/download, accessed on 28 June 2023. For this study, the yearly history for 2019 was employed. The yearly history combines two water classes, seasonal and permanent. The UA and PA for the entire time series, including the 2019 map, exceed 95%.

The FNF map is a thematic LC map that classifies forested regions worldwide [29]. Developed by the Japan Aerospace Exploration Agency (JAXA), it provides a multi-temporal representation of forest areas with irregular time intervals. The map covers the periods from 2007 to 2010 and from 2015 to 2020, categorizing areas as forest, water, or not water, with an approximate resolution of 25 m. The FNF map for 2019 was used in this research. The accuracy of this specific product is not specified. The product is distributed in two tile sizes: 1° × 1° or 5° × 5° from https://www.eorc.jaxa.jp/ALOS/index_e.htm, accessed on 28 June 2023.

MapBiomas project focuses on generating maps for six Brazilian biomes, namely the Amazon, Atlantic Forest, Cerrado, Caatinga, Pampa, and Pantanal [53]. These maps provide a general overview of LC types at 30 m of spatial resolution annually, dating back to 1985. MapBiomas utilizes a hierarchical legend with three levels of classification. The first level consists of six broad classes, which are further subdivided into more specific classes at the second and third levels. Since its inception in 2016, the project has undergone several collections with different data processing methodologies. The MOLCA creation specifically used the map from Collection 7 for the year 2019, which achieved an OA of 89% [54]. MapBiomas is licensed under the Creative Commons CC-BY-SA license, which means it is freely available and can be accessed through various means, including GoogleEarthEngine (GEE), the GEE app—Toolkit, the MapBiomas dashboard, the QGIS plugin, or direct download in GeoTiff format via a provided link on

https://mapbiomas.org/en/colecoes-mapbiomas-1?cama_set_language=en, accessed on 28 June 2023. The default CRS used is WGS84.

The CCI Africa Prototype is a general-purpose LC map with a resolution of 20 m, produced by the ESA CCI LC team, representing the LC state in Africa for the year 2016. The legend of the CCI Africa Prototype consists of Tree-covered areas, Shrub-covered areas, Grassland, Cropland, Vegetation aquatic or regularly flooded, Lichen and mosses/sparse vegetation, Bare areas, Built up areas, and Snow and/or ice and open water. The product can be downloaded as a single GeoTiff file in WGS84 CRS for the entire African continent from https://2016africalandcover20m.esrin.esa.int, accessed on 28 June 2023. Accuracy assessments of the CCI Africa Prototype were conducted for four countries: Kenya, Gabon, Ivory Coast, and South Africa [55]. The OA was found to be 44% for South Africa, 47% for Ivory Coast, 56% for Kenya, and 91% for Gabon.

The ESA DUE GlobPermafrost map describes the LC of permafrost regions, including Western Siberia (Russia), Barrow (Alaska), Teshekpuk (Alaska), Mackenzie Delta (Canada), Umiuaq (Canada), Kytalyk (Russia), Lena Delta (Russia), Seward Peninsula (Alaska), and Yukon Delta (Alaska) [56]. The legend of ESA DUE GlobPermafrost is very detailed on polar LC types (21 in total). Each permafrost region has a corresponding GeoTiff file, which can be downloaded from the PANGAEA data publisher under the Creative Commons Attribution 4.0 International license [57]. The CRS used for ESA DUE GlobPermafrost is the UTM projection, and the OA of the map is estimated to be 83%.

2.2. MOLCA Methodology Concepts

The creation of MOLCA involves intersecting multiple HRLC datasets to determine areas of agreement. Only the areas where all the HRLCs agree are retained, while pixels showing LC class discrepancies among the intersected HRLCs are designated as null. From a theoretical standpoint, there is a high probability that the MOLCA has high accuracy, because a manyfold agreement increases the odds of pixels being accurate [58]. Pixels that are accurately classified have a high likelihood of being found in corresponding positions across different datasets, as correct classification is a primary objective during the classification process. Conversely, errors in the LC derivation result from undesired factors associated with the classification process. These errors can be influenced by various factors, such as the quality and quantity of training data, the suitability of the classification algorithm, the accuracy and quality of satellite imagery, the complexity of the LC types being classified, as well as the presence of atmospheric phenomena such as clouds. Since different agencies and procedures are responsible for producing most of the existing HRLCs, it is expected that errors in different datasets are independent and not replicated across them. Thus, the MOLCA methodology's primary benefit arises from its utilization of multiple HRLC maps to create the training dataset. This approach is anticipated to improve the classification accuracy compared to relying solely on training samples extracted from individual HRLC-existing maps.

2.3. MOLCA Generation Procedure

A schema of the MOLCA generation procedure is shown in Figure 2. Different parts of the procedure are grouped into preparation, data harmonization, and MOLCA generation.

The procedure of creating MOLCA started by downloading the identified HRLC products (see Table 1) for regions of interest. This was conducted automatically with Python [59] when feasible, otherwise, it was conducted manually. The legends of these datasets were carefully compared to determine common classes across them. The classes that consistently appeared across multiple datasets were chosen as the target classes for MOLCA. Details on MOLCA legend are included in Table A1 of Appendix A. To align the legends of the existing datasets with the target legend of MOLCA, a correspondence table was created for each dataset and stored in a textual file. Based on correspondence tables, txt files with reclassification rules were created to be used in later steps. The reclassification rules contain information about the original raster value, the target raster value, and the

target class label. The legend harmonization was performed manually because a single class might have a different name and code in different HRLCs.

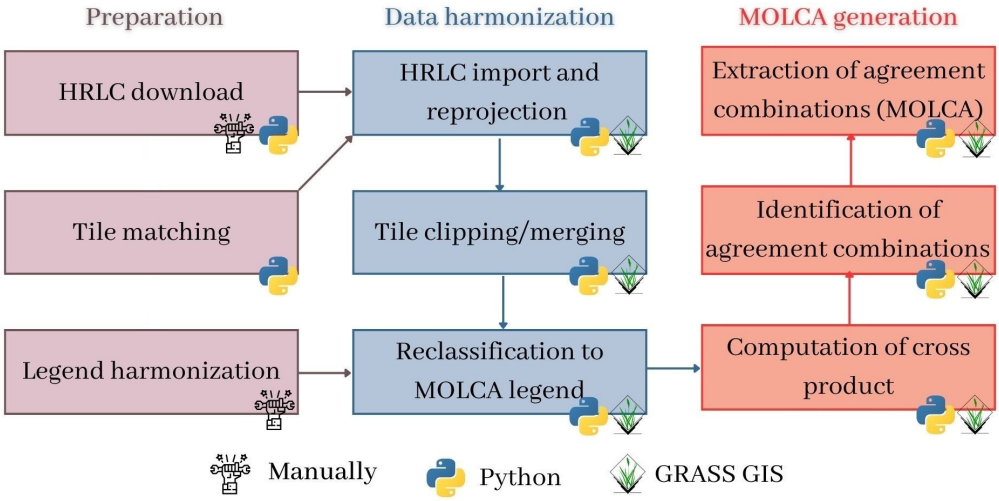

Figure 2. Schema of the MOLCA generation procedure.

Since different datasets used different tiling systems, it was necessary to find matching tiles across the datasets. This matching process was automated using Python pandas, shapely, rasterio, and geopandas libraries. The rest of the procedure was automatized by using a combination of GRASS GIS and Python. Principal GRASS GIS modules used for MOLCA generation included *r.import*, *g.region*, *r.stats*, *r.reclass*, *r.patch*, *r.category*, and *r.cross* [60]. These modules were run through the *grass.script* library of Python. Some Python libraries independent of GRASS GIS were also used, such as *numpy*.

A reference dataset was selected to serve as a guidance for the CRS, extent, and spatial resolution of MOLCA tiles. A tile from the ESA CCI HRLC product was chosen, which had a size of 100 km × 100 km, a resolution of 10 m, and used the WGS84 CRS. Information about spatially matching tiles of reference dataset with non-reference datasets was stored in a CSV file.

Prior to importing data into the GRASS GIS database, its default CRS was set to WGS84 CRS. Then, the datasets were imported and automatically reprojected if their source CRS was not WGS84. The extent and resolution of the imported non-reference data tiles were adjusted to the ones of reference tiles. These non-reference tiles were clipped or merged to match the extent of the reference tile. Furthermore, the non-reference tiles were reclassified in accordance with the MOLCA target legend, following reclassification rules. Resampling to target resolution was conducted on the fly when processing operations were executed.

The next step was to extract areas of agreement. A cross-product was computed from the non-reference datasets, which generated a raster map with different values representing combinations of class values found within the input layers. The cross-product labels were analyzed to identify agreement labels, which were defined as labels that appeared consistently across all input HRLCs or at least two HRLCs, with other labels considered null. The agreement labels and their associated values were converted into reclassification rules. Finally, the cross-product was reclassified into the MOLCA.

2.4. MOLCA Validation

The accuracy of MOLCA was evaluated against photo-interpreted samples collected by the authors in one part of the African region. The accuracy metrics were determined using a conventional error matrix [61] which was filled with classes derived from photo-interpretation, along with their corresponding MOLCA classes found at the same sampling locations. The number of samples was estimated based on Cochran's equation [62]. The sample count was determined to be 1068; an additional 130 samples were preven-

tively included to take into account the chance of discarding some samples due to photo-interpretation uncertainties.

Each class within the MOLCA was considered a distinct stratum. An equal number of samples was selected in each stratum, with the exception of Bareland and Wetland because their count in MOLCA was low. Consequently, the number of samples for these classes was set to match the maximum number of pixels present in MOLCA, specifically 22 for Bareland and 6 for Wetland. The samples within each stratum, except for Bareland and Wetland, were randomly selected, while all pixels belonging to Bareland and Wetland were converted into samples.

The sampling survey was designed in Open Foris Collect platform [63], while photo interpretation was conducted in Open Foris Collect Earth software [63] where the photo-interpreter could use either Google imagery or temporal profiles of vegetation indices from Landsat 7/8, Sentinel-2, and MODIS imagery to assign a class label to a sample. A total of 148 samples were discarded due to low confidence in the photo-interpretation deriving from poor image quality, clouds, or a high degree of similarity with other classes. The remaining 1050 samples were used for MOLCA validation.

3. Results

A total of 2075 MOLCA tiles were created and made publicly available on Zenodo (https://doi.org/10.5281/zenodo.8071675, accessed on 28 June 2023). The datasets include 893 tiles in the African, 658 tiles in the Amazonian, and 524 tiles in the Siberian macro-region of interest. The tiling grid used for MOLCA follows the Sentinel-2 Level-1C product tiling grid. Accordingly, the identifier assigned to each MOLCA tile corresponds to the identifier of the corresponding Sentinel-2 Level-1C tile. Figure 3 shows an example of MOLCA tile in the Amazon region.

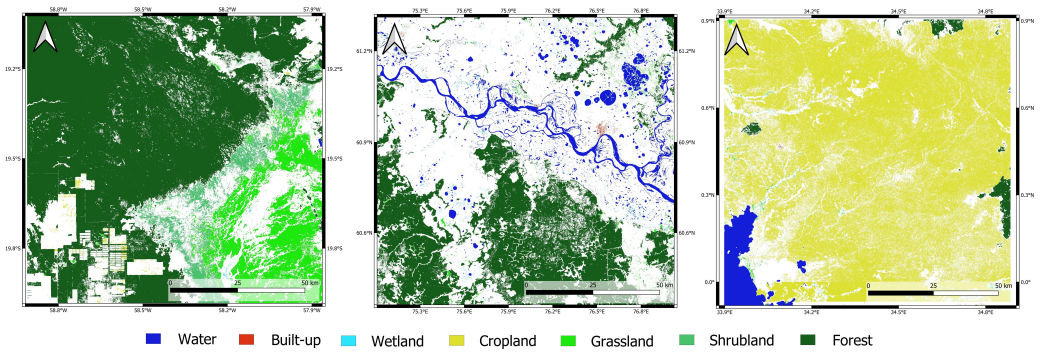

Figure 3. Example of MOLCA tile in the Amazon region (**left**), in Siberian region (**center**), and African region (**right**). The names of the tiles from left to right in order are *MOLCA_21KUU_v1.tif*, *MOLCA_43VEH_v1.tif*, *MOLCA_36NXF_v1.tif*, where the identifiers are 21KUU, 43VEH, and 36NXF.

In addition to the tile data, two supplementary files are also provided. The first file contains the vector representation of MOLCA tile extents along with statistics such as the number of pixels per class, the total number of pixels, and the proportion of valid values for each tile in the attribute table. The second file in the CSV format includes the class codes and labels for MOLCA.

3.1. MOLCA Statistics

Table 2 presents an overview of MOLCA statistics in terms of the class-wise number of pixels and the number of HRLCs participating in the generation of each class in each region, as well as the total number of pixels and the proportion of MOLCA in each region of interest regardless of class.

Table 2. Statistics of MOLCA: number of pixels extracted for each class and the total number of pixels per region, the proportion of the MOLCA in a region of interest, and number of existing HRLCs that participated in the creation of a specific class. The number of HRLCs participating in the generation of MOLCA in Siberia is denoted by "*" because ESA DUE GlobPermafrost is not covering the entire region. "#" stands for "number of".

	Siberia		**Amazon**		**Africa**	
	# Pixels	**# Maps**	**# Pixels**	**# Maps**	**# Pixels**	**# Maps**
Bareland	70,522,207	3 *	10,974,208	3	17,527,789,276	3
Built-up	11,954,315	4	64,772,162	5	10,054,752	5
Cropland	3,045,996,831	2	4,740,455,996	3	4,142,663,163	3
Forest	15,748,595,107	4 *	26,141,725,251	4	13,429,492,002	4
Grassland	5,725,978,494	3 *	5,468,110,102	3	4,493,491,684	3
Permanent ice and snow	78,840,342	2	0		0	
Shrubland	1,763,096	3 *	4,109,823,259	3	2,174,109,509	3
Water	5,424,855,889	5 *	1,718,120,337	5	2,550,708,631	5
Wetland	393,196,640	3 *	82,520,517	3	5,369,604	3
Total # pixels	30,501,702,921		42,336,501,832		44,206,814,559	
Proportion of MOLCA in region of interest	43%		52%		40%	

3.2. Accuracy

The result of the validation procedure based on 1050 samples is represented by the error matrix and accuracy indexes of MOLCA, reported in Table 3. The accuracy indexes include UA, PA, F1 score, OA, Kappa, and False Discovery Rate (FDR) The rows of the error matrix represent MOLCA classes, while the columns represent classes of photo-interpreted reference samples and accuracy indexes.

Table 3. Accuracy of MOLCA in the African region of interest.

	Bareland	**Built-Up**	**Cropland**	**Forest**	**Grassland**	**Shrubland**	**Water**
Bareland	0	0	0	0	0	0	0
Built-up	0	210	0	0	0	0	0
Cropland	0	3	76	5	18	2	0
Forest	0	0	0	184	0	0	0
Grassland	3	4	4	2	158	3	0
Shrubland	0	0	0	0	1	191	0
Water	0	0	0	0	0	0	186
UA	0%	100%	73%	100%	91%	99%	100%
PA	0%	97%	95%	96%	89%	97%	100%
F1 score	0%	98%	83%	98%	90%	98%	100%
OA				96%			
Kappa				95%			
FDR				4%			

4. Discussion

MOLCA dataset has billions of LC pixels for the three regions of interest. Its legend and temporal representatives are determined based on the characteristics of input HRLCs.

To be included in MOLCA, a class must appear in at least two input HRLCs. Additionally, the other input HRLCs should either have the same class or no class at all. If these conditions are not met, the class will not be part of MOLCA. Moreover, the intersection procedure eliminates small differences in legends between input HRLC datasets. When there is a variation in the definition of a specific class between different HRLCs, the MOLCA derivation procedure ensures that only the common characteristics are retained. To illustrate, if one HRLC defines Forest as an area of trees with at least 2 m height, while another HRLC sets the threshold at 5 m, MOLCA will exclude any Forest patches with trees shorter than 5 m during the intersection procedure. This exclusion occurs because there is no agreement on the representation of Forest with trees below 5m in the second dataset. If classes with the same name are significantly different in their definition, it might happen that they do not constitute any agreement during MOLCA derivation, and therefore they will be eliminated.

Similarly, if there is a difference in the baseline years of input HRLCs, and a land cover change happened between these years if the HRLC with a more recent baseline year captures the changes, it will cause disagreement among the HRLCs for pixels affected by the change, and consequently, such pixels will not be present in MOLCA dataset. Hence, MOLCA's temporal representativeness falls between the most recent and the least recent baseline year of the input HRLCs.

By combining FROM-GLC, GL30, WSF, GSW, FNF, GHS BU S1NODSM, Mapiomas (Amazon only), CCI Africa Prototype (Africa only), and ESA DUE GlobPermafrost (Siberia only) MOLCA's legend resulted in Bareland, Built-up, Cropland, Forest, Grassland, Shrubland, Water, and Wetland classes in all regions, plus the Permanent ice and snow class in Siberia. MOLCA legend and its correspondence to the Food and Agriculture Organization (FAO) Land Cover Classification System (LCCS) is displayed in Table 4. The MOLCA legend aligns with the second out of three levels of the dichotomous phase of the FAO LCCS. In the first level of FAO LCCS, classes are distinguished based on the presence of vegetation, categorized as (A) primarily vegetated and (B)primarily non-vegetated. The second level further discriminates based on the presence of water, distinguishing between (1) terrestrial and (2) aquatic. The third level considers the artificiality of LC. In MOLCA, the vegetation classes are not fully differentiated as per the third level of FAO LCCS, i.e., most classes are not discriminated by artificiality. This drawback hampers the effectiveness of MOLCA, as FAO LCCS is currently the solely available system that facilitates the interoperability of legends of different LC datasets through a hierarchical approach. However, the inherent nature of MOLCA restricts the control over the legend. Despite this, it is worth noting that the legend remains compatible with the majority of existing HRLCs, which is a de facto standard legend.

Since not all classes are derived from the same HRLCs (e.g., water is present in five input HRLCs, while Grassland is present in three of them), the temporal representatives of each class vary. Nonetheless, MOLCA provides an approximate but reliable representation of land cover during the timeframe of 2016–2020, as explained above. Details about temporal representatives of each class in each region are included in Appendix B (see Tables A2–A4).

MOLCA statistics (see Table 2) show that the Forest class emerges as the most abundant class within MOLCA. Among the HRLCs employed, the Built-up and Water classes have the highest representation with five HRLCs, followed by the Forest class with four HRLCs. Other classes primarily rely on three HRLCs, except for Cropland and Permanent ice and snow in Siberia.

Accuracy results (see Table 3) indicate a general high accuracy given that that the OA of MOLCA is 96%, Kappa index is 95%, and FDR is 4%. Regarding the classes, UA and PA scores exceed 85%, except for the Cropland class. Cropland has a UA of 73%; thus, it exhibits moderate overestimation. Unfortunately, no confident samples were available for the Wetland class, making it impossible to estimate its accuracy. It should be noted that the Bareland class had only three samples, which may not accurately reflect its classification accuracy. F1 score is very high for the majority of classes (>90%), which indicates high

accuracy of classes, and a good balance between UA and PA. Being derived from UA and PA, the F1 score is slightly lower for the Cropland class (i.e., 80%) than for other classes, and equal to 0% in the case of Bareland class, for the above-mentioned reasons.

One limitation of MOLCA is the lack of pixels of Permanent ice and snow in Africa. It is present in high mountain peaks, but it is not identified in MOLCA. There are two possible reasons for this issue. Firstly, the area of the class is extremely small in the African region of interest; it significantly reduced the possibility of existing HRLCs having a consensus on such a narrow area. Secondly, it could be that the class is not detected by one of the input HRLCs, and consequently, consensus among HRLCs was not possible.

Table 4. Legend of MOLCA and its correspondence to FAO LCCS.

MOLCA Code	MOLCA Class	LCCS Code	LCCS Description
8	Cropland	A11	Primarily non-vegetated, Terrestrial, Bare areas
20	Forest	A12, A11	Primarily vegetated, Terrestrial, Cultivated and managed areas
7	Grassland	A12, A11	Primarily vegetated, Terrestrial, Semi-natural vegetation, and Cultivated and managed areas
5	Shrubland	A12, A11	Primarily vegetated, Terrestrial, Semi-natural vegetation, and Cultivated and managed areas
9	Wetland	A24, A22	Primarily non-vegetated, Terrestrial, Artificial surfaces
13	Built-up	B15	Primarily vegetated, Terrestrial, Semi-natural vegetation, and Cultivated and managed areas
12	Bareland	B16	Primarily non-vegetated, Aquatic or regularly flooded, Natural waterbodies, snow and ice, and Artificial waterbodies, snow and ice
16	Permanent ice and snow	B28, B27	Primarily non-vegetated, Aquatic or regularly flooded, Natural waterbodies, snow and ice, and Artificial waterbodies, snow and ice
15	Water	B28, B27	Primarily vegetated, Aquatic or regularly flooded, Semi-natural vegetation, and Cultivated and managed areas

5. Conclusions and Outlook

The overall objective of MOLCA is to establish a benchmark framework for deriving training datasets from existing HRLC datasets. The HRLCs are combined by the intersection method which ensures that only regions where all datasets align in terms of classes are retained, while conflicting areas are eliminated. Such a manyfold agreement increases the probability that MOLCA retained only correct portions of input HRLCs. Currently, MOLCA covers three macro-regions of the world, two of which are in Siberia and Africa which are rarely included in existing training benchmark datasets [17,42–49]. Another advantage of MOLCA is that it provides 117 billion of 10-m pixels, or 43% of coverage of the region of interest, which is, to our best knowledge, significantly more compared to any other existing training benchmark datasets. Such a large number of pixels is suitable to support deep learning techniques that are gaining popularity and require extensive training datasets. Nevertheless, it can also support traditional ML approaches.

The results of the accuracy evaluation demonstrated an OA of 96%, Kappa index equals to 95%, and low FDR (i.e., 4%), all of which indicate very high accuracy. Among the seven categories evaluated, four of them exhibited an accuracy rate surpassing 90% for both UA and PA. The Grassland category demonstrated UA and PA values nearing 90%. On the other hand, UA for the Cropland category implies a potential overestimation of the Cropland class. The F1 score for each class was in line with UA and PA results. While MOLCA may not achieve perfect accuracy for certain classes, a study by Rolnick et al. [64] shows that the noisy training samples do not significantly affect the performance of deep neural networks as long as the training dataset is sufficiently large. Moreover, some of the currently available HRLCs were based on other LC products, and in some cases without

taking into account that each LC product contains some degree of error. Therefore, we argue that MOLCA might be more suitable for training data than an individual HRLC dataset.

The MOLCA legend is similar to most of the worldwide HRLCs and consists of different types of LC such as Bareland, Built-up, Cropland, Forest, Grassland, Shrubland, Water, Wetland, and Permanent ice and snow (in Siberia only). Although it does not fully align with FAO LCCS, it shows promise in aiding HRLC production if the current legend trend continues.

As a future development of this work, we plan to incorporate other recently published global HRLCs into the MOLCA derivation procedure, since they were not available at the time of generation of this version. On one hand, this would be useful for further refining MOLCA and increasing its accuracy, and on the other hand, it would allow the exploration of a suitable combination of existing HRLCs to ensure the representation of extremely small classes in MOLCA that currently are an issue (e.g., Permanent ice and snow in Africa). Furthermore, we also plan to expand MOLCA availability to other regions of the world.

Author Contributions: Conceptualization, G.B., M.A.B. and D.O.; methodology, G.B.; software, G.B.; validation, G.B.; formal analysis, G.B.; investigation, G.B.; data curation, G.B.; writing—original draft, G.B.; writing—review and editing, G.B. and D.O.; visualization, G.B.; supervision, M.A.B. and D.O.; project administration, M.A.B.; funding acquisition, M.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the European Space Agency (ESA) within the project CCI+ HRLC—Climate Change Initiative Extension (CCI+), Phase 1: New Essential Climate Variables (NEW ECVS) High Resolution Land Cover ECV (HR_LandCover_cci) (https://climate.esa.int/en/projects/high-resolution-land-cover, accessed on 28 June 2023).

Data Availability Statement: The resulting datasets of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.8071675 (accessed on 28 June 2023) under a CC BY 4.0 license. Raw data used to produce the resulting datasets are available in the public domain.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CCI	Climate Change Initiative
CRS	Coordinate Reference System
DLR	German Aerospace Center
DUE	Data User Element
ESA	European Space Agency
FAO	Food and Agriculture Organization
FDR	False Discovery Rate
FNF	Forest Non-Forest
FROM-GLC	Finer Resolution Observation and Monitoring of Global Land Cover
GEE	Google Earth Engine
GHS BU	Global Human Settlements—Built-Up
GHS BU S1NODSM	Global Human Settlements—Built-Up Sentinel-1-derived
GMW	Global Mangrove Watch
GSW	Global Surface Water
HRLC	High-Resolution Land Cover
IUCN	International Union for Conservation of Nature
JAXA	Japan Aerospace Exploration Agency
JRC	Joint Research Centre
LC	Land Cover
LCCS	Land Cover Classification System

LRLC	Low-Resolution Land Cover
ML	Machine Learning
MOLCA	Map Of Land Cover Agreement
MRLC	Medium-Resolution Land Cover
NGCC	National Geomatics Center of China
OA	Overall Accuracy
PA	Producer's Accuracy
UA	User's Accuracy
UTM	Universal Transverse Mercator
WGS84	World Geodetic System 1984
WSF	World Settlement Footprint

Appendix A

MOLCA legend was based on the classes that consistently showed up in various datasets used for MOLCA creation. One of the initial steps in MOLCA creation was to find matching classes across different datasets. It resulted in a correspondence table that shows the resulting MOLCA legend, and how classes of other HRLCs correspond to this legend (see Table A1).

Table A1. Table of correspondence of legends of different HRLCs.

MOLCA	FROM-GLC	GL30	GHS BU S1-NODSM	WSF	GSW	FNF	CCI Africa Prototype	ESA DUE GlobPermafrost	MapBiomas
Bareland	Bare land	Bare land					Bare areas	Sparse vegetation (without shrubs), mostly sandy soil, flood plains, recent landslides, also within fire scars; Barren, rare vegetation (petrophytes and psammophytes)	Salt flat; Rocky outcrop; Beach; Dune and sand spot; Mining; Other non vegetated areas
Cropland	Cropland	Cultivated land					Cropland		Agriculture; Temporary crop; Mosaic of uses
Forest	Forest	Forest				Forest	Trees cover areas	Tall shrubs, deciduous forest; Coniferous (partially mixed) forest	Forest formation; Forest plantation
Grassland	Grass	Grassland					Grassland	Meadows, grass and herb-dominated	Grassland; Pasture
Built-up	Impervious	Artificial surfaces	Built-up	Settlements			Built up areas		Urban area
Shrubland	Shrub	Shrubland					Shrubs cover areas	Graminoid, prostrate dwarf shrub, patterned ground, partially bare; Dry to moist prostrate to erect dwarf shrub tundra; Moist to wet graminoid prostrate to erect dwarf shrub tundra; Wet to waterlogged graminoid prostrate to low shrub tundra; Moist low dense shrubs	Savanna formation

Table A1. *Cont.*

MOLCA	FROM-GLC	GL30	GHS BU S1-NODSM	WSF	GSW	FNF	CCI Africa Prototype	ESA DUE GlobPermafrost	MapBiomas
Permanent ice and snow	Snow/Ice	Permanent snow and ice					Snow and/or ice		
Water	Water	Water bodies			Seasonal water; Permanent water		Open water	Floodplain, mostly fluvial; Seasonally inundated, Water (shallow or high sediment yield); Water (medium depth or medium sediment yield); Water (low sediment yield)	River; Lake and ocean; Aquaculture
Wetland	Wetland	Wetland					Vegetation aquatic or regularly flooded	Floodplain, mostly lacustrine	Mangrove; Wetland

Appendix B

Table A2 is an L-shaped diagram that displays the relationship between the HRLC datasets (rows) and the unique LC classes (columns) represented in those datasets. If a particular HRLC dataset includes a particular LC class, the corresponding cell in the table is highlighted. The table also includes information about the baseline year for each HRLC dataset in Siberia. The MOLCA represents the LC in a specific time period, which is determined by the minimum and maximum baseline years of the datasets used to create it. The period of representativeness for each class may vary because the classes are not derived from the same datasets.

For the MOLCA in Siberia, the Cropland, Permanent ice and snow, and Shrubland classes are representative of the LC in 2017, while the Bareland, Forest, Grassland, and Wetland classes are representative of the period between 2016 and 2017. The Built-up and Water classes are representative of the period between 2016 and 2019.

Table A2. L-diagram of existing HRLCs (rows) and their classes (columns) in Siberia.

Existing HRLCs in Siberia	Year	Bareland	Built-Up	Cropland	Forest	Grassland	Permanent	Ice and Snow	Shrubland	Water	Wetland
FROM-GLC	**2017**	■	■	■	■	■	■	■	■	■	■
GL30	**2017**	■	■	■	■	■	■	■	■	■	■
ESA DUE GlobPermafrost	**2016**	■			■	■			■	■	■
GHS BU S1NODSM	**2016**		■								
WSF	**2019**		■								
FNF	**2017**				■					■	
GSW	**2019**									■	

The L-shaped diagram for the Amazon region of interest is displayed in Table A3, and the one for Africa in Table A4. For the MOLCA in Amazon, classes Bareland, Cropland, Forest, Grassland, Shrubland, Water, and Wetland are representative for 2017–2019, and class Built-up for the period 2016–2019. In the case of Africa, classes Bareland, Cropland, Forest, Grassland, Shrubland, and Wetland are representative for the period 2016–2017, and classes Water and Built-up for the period 2016–2019.

Table A3. L-diagram of existing HRLCs (rows) and their classes (columns) in the Amazon.

Existing HRLCs in Amazon	Year	Bareland	Built-Up	Cropland	Forest	Grassland	Shrubland	Water	Wetland
FROM-GLC	**2017**	■	■	■	■	■	■	■	■
GL30	**2017**	■	■	■	■	■	■	■	■
MapBiomas	**2019**	■	■	■	■	■	■	■	■
GHS BU S1NODSM	**2016**		■						
WSF	**2019**		■						
FNF	**2017**				■			■	
GSW	**2019**							■	

Table A4. L-diagram of existing HRLCs (rows) and their classes (columns) in Africa.

Existing HRLCs in Africa	Year	Bareland	Built-Up	Cropland	Forest	Grassland	Shrubland	Water	Wetland
FROM-GLC	**2017**	■	■	■	■	■	■	■	■
GL30	**2017**	■	■	■	■	■	■	■	■
CCI Africa Prototype	**2016**	■	■	■	■	■	■	■	■
GHS BU S1NODSM	**2016**		■						
WSF	**2019**		■						
FNF	**2017**				■			■	
GSW	**2019**							■	

References

1. Bontemps, S.; Defourny, P.; Radoux, J.; Van Bogaert, E.; Lamarche, C.; Achard, F.; Mayaux, P.; Boettcher, M.; Brockmann, C.; Kirches, G.; et al. Consistent Global Land Cover Maps for Climate Modelling Communities: Current Achievements of the ESA' Land Cover CCI. 2013. Available online: https://ui.adsabs.harvard.edu/abs/2013ESASP.722E..62B (accessed on 26 May 2023).
2. Haddad, N.M.; Brudvig, L.A.; Clobert, J.; Davies, K.F.; Gonzalez, A.; Holt, R.D.; Lovejoy, T.E.; Sexton, J.O.; Austin, M.P.; Collins, C.D.; et al. Habitat Fragmentation and its Lasting Impact on Earth'S Ecosystems. *Sci. Adv.* **2015**, *1*, e1500052. [CrossRef] [PubMed]
3. Abe, C.; Lobo, F.; Dibike, Y.; Costa, M.; Dos Santos, V.; Novo, E. Modelling the Effects of Historical and Future Land Cover Changes on the Hydrology of an Amazonian Basin. *Water* **2018**, *10*, 932. [CrossRef]
4. Birhanu, A.; Masih, I.; van der Zaag, P.; Nyssen, J.; Cai, X. Impacts of Land Use and Land Cover Changes on Hydrology of the Gumara Catchment, Ethiopia. *Phys. Chem. Earth Parts A/B/C* **2019**, *112*, 165–174. [CrossRef]
5. Cui, G.; Lee, W.K.; Kwak, D.A.; Choi, S.; Park, T.; Lee, J. Desertification Monitoring by LANDSAT TM Satellite Imagery. *For. Sci. Technol.* **2011**, *7*, 110–116. [CrossRef]
6. Enoguanbhor, E.; Gollnow, F.; Nielsen, J.; Lakes, T.; Walker, B. Land Cover Change in the Abuja City-Region, Nigeria: Integrating GIS and Remotely Sensed Data to Support Land Use Planning. *Sustainability* **2019**, *11*, 1313. [CrossRef]
7. Picoli, M.C.A.; Camara, G.; Sanches, I.; Simões, R.; Carvalho, A.; Maciel, A.; Coutinho, A.; Esquerdo, J.; Antunes, J.; Begotti, R.A.; et al. Big Earth Observation Time Series Analysis for Monitoring Brazilian Agriculture. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 328–339. [CrossRef]
8. Zhu, Z.; Wulder, M.A.; Roy, D.P.; Woodcock, C.E.; Hansen, M.C.; Radeloff, V.C.; Healey, S.P.; Schaaf, C.; Hostert, P.; Strobl, P.; et al. Benefits of the Free and Open Landsat Data Policy. *Remote Sens. Environ.* **2019**, *224*, 382–385. [CrossRef]
9. Woodcock, C.E.; Allen, R.; Anderson, M.; Belward, A.; Bindschadler, R.; Cohen, W.; Gao, F.; Goward, S.N.; Helder, D.; Helmer, E.; et al. Free Access to Landsat Imagery. *Science* **2008**, *320*, 1011. [CrossRef]
10. Copernicus Open Access Hub. Legal Notice on the Use of Copernicus Sentinel Data and Service Information. Available online: https://sentinels.copernicus.eu/documents/247904/690755/Sentinel_Data_Legal_Notice (accessed on 31 May 2023).
11. Burke, M.; Driscoll, A.; Lobell, D.B.; Ermon, S. Using Satellite Imagery to Understand and Promote Sustainable Development. *Science* **2021**, *371*, eabe8628. [CrossRef]
12. Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; Aroyo, L.M. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; ACM: Yokohama, Japan, 2021; pp. 1–15. [CrossRef]
13. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools, and Challenges for the Community. *J. Appl. Remote Sens.* **2017**, *11*, 042609. [CrossRef]
14. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep Learning in Environmental Remote Sensing: Achievements and Challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [CrossRef]
15. Scott, G.J.; England, M.R.; Starms, W.A.; Marcum, R.A.; Davis, C.H. Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 549–553. [CrossRef]
16. Dimitrovski, I.; Kitanovski, I.; Kocev, D.; Simidjievski, N. Current Trends in Deep Learning for Earth Observation: An Open-Source Benchmark Arena for Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 18–35. [CrossRef]
17. Sumbul, G.; De Wall, A.; Kreuziger, T.; Marcelino, F.; Costa, H.; Benevides, P.; Caetano, M.; Demir, B.; Markl, V. BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 174–180. [CrossRef]
18. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]

19. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. MLRSNet: A Multi-Label High Spatial Resolution Remote Sensing Dataset for Semantic Scene Understanding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 337–350. [CrossRef]
20. Bratic, G.; Vavassori, A.; Brovelli, M.A. Review of High-Resolution Global Land Cover. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 175–182. [CrossRef]
21. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S.; et al. Dynamic World, Near Real-Time Global 10 m Land Use Land Cover Mapping. *Sci. Data* **2022**, *9*, 251. [CrossRef]
22. Karra, K.; Kontgis, C.; Statman-Weil, Z.; Mazzariello, J.C.; Mathis, M.; Brumby, S.P. Global Land Use/Land Cover with Sentinel 2 and Deep Learning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; IEEE: Brussels, Belgium, 2021; pp. 4704–4707. [CrossRef]
23. Gong, P.; Wang, J.; Yu, L.; Zhao, Y.; Zhao, Y.; Liang, L.; Niu, Z.; Huang, X.; Fu, H.; Liu, S.; et al. Finer Resolution Observation and Monitoring of Global Land Cover: First Mapping Results with Landsat TM and ETM+ Data. *Int. J. Remote Sens.* **2013**, *34*, 2607–2654. [CrossRef]
24. Li, C.; Gong, P.; Wang, J.; Zhu, Z.; Biging, G.S.; Yuan, C.; Hu, T.; Zhang, H.; Wang, Q.; Li, X.; et al. The First All-Season Sample Set for Mapping Global Land Cover with Landsat-8 Data. *Sci. Bull.* **2017**, *62*, 508–515. [CrossRef]
25. Gong, P.; Liu, H.; Zhang, M.; Li, C.; Wang, J.; Huang, H.; Clinton, N.; Ji, L.; Li, W.; Bai, Y.; et al. Stable Classification with Limited Sample: Transferring a 30-M Resolution Sample Set Collected in 2015 to Mapping 10-M Resolution Global Land Cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. [CrossRef] [PubMed]
26. Marconcini, M.; Metz-Marconcini, A.; Üreyen, S.; Palacios-Lopez, D.; Hanke, W.; Bachofer, F.; Zeidler, J.; Esch, T.; Gorelick, N.; Kakarla, A.; et al. Outlining Where Humans Live—The World Settlement Footprint 2015. *Sci. Data* **2020**, *7*, 242. [CrossRef] [PubMed]
27. Marconcini, M.; Marconcini, A.M.; Esch, T.; Gorelick, N. Understanding Current Trends in Global Urbanisation—The World Settlement Footprint Suite. *GI_Forum* **2021**, *9*, 33–38. [CrossRef]
28. Pekel, J.F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution Mapping of Global Surface Water and its Long-Term Changes. *Nature* **2016**, *540*, 418–422. [CrossRef] [PubMed]
29. Shimada, M.; Itoh, T.; Motooka, T.; Watanabe, M.; Shiraishi, T.; Thapa, R.; Lucas, R. New Global Forest/Non-Forest Maps from ALOS PALSAR Data (2007–2010). *Remote Sens. Environ.* **2014**, *155*, 13–31. [CrossRef]
30. Chen, J.; Chen, J.; Liao, A.; Cao, X.; Chen, L.; Chen, X.; He, C.; Han, G.; Peng, S.; Lu, M.; et al. Global Land Cover Mapping at 30M Resolution: A Pok-Based Operational Approach. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 7–27. [CrossRef]
31. Pesaresi, M.; Corbane, C.; Julea, A.; Florczyk, A.J.; Syrris, V.; Soille, P. Assessment of the Added-Value of Sentinel-2 for Detecting Built-Up Areas. *Remote Sens.* **2016**, *8*, 299. [CrossRef]
32. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21St-Century Forest Cover Change. *Science* **2013**, *342*, 850–853. [CrossRef]
33. Bunting, P.; Rosenqvist, A.; Lucas, R.; Rebelo, L.M.; Hilarides, L.; Thomas, N.; Hardy, A.; Itoh, T.; Shimada, M.; Finlayson, C. The Global Mangrove Watch—A New 2010 Global Baseline of Mangrove Extent. *Remote Sens.* **2018**, *10*, 1669. [CrossRef]
34. Bunting, P.; Rosenqvist, A.; Hilarides, L.; Lucas, R.M.; Thomas, N.; Tadono, T.; Worthington, T.A.; Spalding, M.; Murray, N.J.; Rebelo, L.M. Global Mangrove Extent Change 1996–2020: Global Mangrove Watch Version 3.0. *Remote Sens.* **2022**, *14*, 3657. [CrossRef]
35. Van De Kerchove, R.; Zanaga, D.; De Keersmaecker, W.; Li, L.; Tsendbazar, N.; Lesiv, M.; Arino, O. World Cover: Product User Manual. 2020. Available online: https://esa-worldcover.s3.amazonaws.com/v100/2020/docs/WorldCover_PUM_V1.0.pdf (accessed on 1 June 2023).
36. Potapov, P.; Turubanova, S.; Hansen, M.C.; Tyukavina, A.; Zalles, V.; Khan, A.; Song, X.P.; Pickens, A.; Shen, Q.; Cortez, J. Global Maps of Cropland Extent and Change Show Accelerated Cropland Expansion in the Twenty-First Century. *Nat. Food* **2022**, *3*, 19–28. [CrossRef] [PubMed]
37. Zhang, X.; Liu, L.; Chen, X.; Gao, Y.; Xie, S.; Mi, J. GLC_FCS30: Global Land-Cover Product with Fine Classification System at 30 m Using Time-Series Landsat Imagery. *Earth Syst. Sci. Data* **2021**, *13*, 2753–2776. [CrossRef]
38. National Geomatics Center of China. GlobeLand30: Product Introduction. Available online: http://globeland30.org/Page/EN_sysFrame/dataIntroduce.html?columnID=81&head=product&para=product&type=data (accessed on 12 June 2023).
39. Esri and Microsoft and Impact Observatory. Sentinel-2 Land Use Land Cover Downloader. Available online: https://www.arcgis.com/apps/instant/media/index.html?appid=fc92d38533d440078f17678ebc20e8e2 (accessed on 12 June 2023).
40. Oxoli, D.; Bratic, G.; Wu, H.; Brovelli, M. Extending Accuracy Assessment Procedures of Global Coverage Land Cover Maps Through Spatial Association Analysis. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 1601–1607. [CrossRef]
41. Wang, X.; Chen, L.; Ban, T.; Lyu, D.; Guan, Y.; Wu, X.; Zhou, X.; Chen, H. Accurate Label Refinement from Multiannotator of Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
42. Chaudhuri, B.; Demir, B.; Chaudhuri, S.; Bruzzone, L. Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 1144–1158. [CrossRef]
43. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [CrossRef]

44. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
45. Zhao, B.; Zhong, Y.; Xia, G.S.; Zhang, L. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2108–2123. [CrossRef]
46. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
47. Zhu, X.X.; Hu, J.; Qiu, C.; Shi, Y.; Kang, J.; Mou, L.; Bagheri, H.; Haberle, M.; Hua, Y.; Huang, R.; et al. So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 76–89. [CrossRef]
48. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; ACM: San Jose, CA, USA, 2010; pp. 270–279. [CrossRef]
49. Planet; SCOON. Planet: Understanding the Amazon from Space. Available online: https://kaggle.com/competitions/planet-understanding-the-amazon-from-space (accessed on 4 June 2023).
50. Corbane, C.; Pesaresi, M.; Politis, P.; Syrris, V.; Florczyk, A.J.; Soille, P.; Maffenini, L.; Burger, A.; Vasilev, V.; Rodriguez, D.; et al. Big Earth Data Analytics on Sentinel-1 and Landsat Imagery in Support to Global Human Settlements Mapping. *Big Earth Data* **2017**, *1*, 118–144. [CrossRef]
51. Pesaresi, M.; Ehrlich, D.; Ferri, S.; Florczyk, A.; Carneiro Freire, S.M.; Halkia, S.; Julea, A.M.; Kemper, T.; Soille, P.; Syrris, V. *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014*; Technical Report EUR 27741; Joint Research Centre: Brussels, Belgium, 2016. [CrossRef]
52. Esch, T.; Brzoska, E.; Dech, S.; Leutner, B.; Palacios-Lopez, D.; Metz-Marconcini, A.; Marconcini, M.; Roth, A.; Zeidler, J. World Settlement Footprint 3D—A First Three-Dimensional Survey of the Global Building Stock. *Remote Sens. Environ.* **2022**, *270*, 112877. [CrossRef]
53. Souza, C.M.; Shimbo, J.Z.; Rosa, M.R.; Parente, L.L.; Alencar, A.A.; Rudorff, B.F.T.; Hasenack, H.; Matsumoto, M.; Ferreira, L.G.; Souza-Filho, P.W.M.; et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sens.* **2020**, *12*, 2735. [CrossRef]
54. Mapbiomas Brasil. Accuracy Statistics. Available online: https://mapbiomas.org/en/accuracy-statistics?cama_set_language=en (accessed on 14 June 2023).
55. Lesiv, M.; See, L.; Mora, B.; Pietsch, S.; Fritz, S.; Bun, H.; Sendabo, S.; Kibuchi, S.; Okemwa, J.; Derrik, O.; et al. Accuracy Assessment of the ESA CCI 20m Land Cover Map: Kenya, Gabon, Ivory Coast and South Africa, 2019. Place: Laxenburg, Austria Publisher: WP-19-009. Available online: https://iiasa.dev.local (accessed on 14 June 2023).
56. Bartsch, A.; Widhalm, B. DUE Globpermafrost Product Documentation: Land Cover Prototype V1.0, 2017. Available online:. (accessed on 14 June 2023). [CrossRef]
57. Bartsch, A.; Widhalm, B.; Pointner, G.; Ermokhina, K.A.; Leibman, M.; Heim, B. Landcover Derived from Sentinel-1 and Sentinel-2 Satellite Data (2015–2018) for Subarctic and Arctic Environments, 2019. [CrossRef]
58. Tuia, D.; Munoz-Mari, J. Learning User's Confidence for Active Learning. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 872–880. [CrossRef]
59. Python Software Foundation. *Python 3.10*; Python Software Foundation: Wilmington, DE, USA, 2022. Available online: https://www.python.org (accessed on 23 June 2023).
60. GRASS Development Team. *Geographic Resources Analysis Support System (GRASS GIS) Software, Version 8.2*; Open Source Geospatial Foundation: Beaverton, DE, USA, 2003–2023. Available online: https://grass.osgeo.org (accessed on 23 June 2023).
61. Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **1997**, *62*, 77–89. [CrossRef]
62. Cochran, W.G. *Sampling Techniques*, 3rd ed.; Wiley Series in Probability and Mathematical Statistics; Wiley: New York, NY, USA, 1977.
63. Open Foris. *Open Foris*; FAO: Rome, Italy, 2022. Available online: https://www.openforis.org (accessed on 23 June 2023).
64. Rolnick, D.; Veit, A.; Belongie, S.; Shavit, N. Deep Learning Is Robust to Massive Label Noise. *arXiv* **2018**, arXiv:1705.10694. [CrossRef]