land

MDPI

Article

Validation and Inter-Comparison of Spaceborne Derived Global and Continental Land Cover Products for the Mediterranean Region: The Case of Thessaly

Ioannis Manakos [1,*], Christina Karakizi [2], Ioannis Gkinis [2] and Konstantinos Karantzalos [2]

[1] Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki 57001, Greece
[2] Remote Sensing Laboratory, National Technical University of Athens, Zographos 15780, Greece; chrkarakizi@central.ntua.gr (C.K.); giannisginis53@gmail.com (I.G.); karank@central.ntua.gr (K.K.)
* Correspondence: imanakos@iti.gr; Tel.: +30-231-125-7760

Academic Editor: Andrew Millington
Received: 17 February 2017; Accepted: 2 May 2017; Published: 12 May 2017

Abstract: Space agencies, international and national organisations and institutions recognize the importance of regularly updated and homogenized land cover information, in the context of both nomenclature and spatial resolution. Moreover, ensuring credibility to the users through validated products with transparent procedures is similarly of great importance. To this end, this study contributes with a systematic accuracy performance evaluation of continental and global land cover layers. Confidence levels during validation and a weighted accuracy assessment were designed and applied. Google Earth imagery were employed to assess the accuracy of three land cover products for the years 2010 and 2012. Results indicate high weighted overall accuracy rates of 89, 90, and 86% for CORINE Land Cover 2012, GIO High Resolution Layers, and Globeland30 datasets, respectively. Moreover, their inter-comparison highlights notable differences especially for classes *Artificial Surfaces* and *Water*. The deviation of specific classes from the general producer's and user's accuracy trends were identified. It is concluded that the different aspects of the employed land cover products can be highlighted more transparently and objectively by integrating confidence levels during the reference data annotation, by employing a stratified sampling based on the several Corine Level-3 subclasses and by applying a weighted overall accuracy procedure.

Keywords: land cover mapping; Globeland30; CORINE; GIO layers; Copernicus service; confidence level; accuracy; spaceborne

1. Introduction

Natural and ecological processes reach out of the human-induced limited space, as delineated by administrative boundaries, demanding standardized products beyond locally generated land cover products, to feed models and scenarios. For example, tele-couplings (e.g., in the form of large area acquisitions or climatic changes) are discussed and measured across the globe for their consequences in land use and local societies along with its value return to the global market [1]. Activities on land and sea are increasingly depended upon frequently updated qualitative land cover products.

Recent advances in data provision frequency and accessibility by the global scientific community, the progress in Earth Observation techniques, and big data handling and processing [2] enabled the generation of numerous Continental and Global Land Cover products (C/GLC) with increasing spatial resolution and frequency. Issues and challenges accompany these developments, mainly in matters of data interpretation and categorization, as well as surface objects' delineation and exact location. Compatibility and interpretation issues are already being treated by working groups, such as the European Environment Information and Observation Network (EIONET) Action Group

on Land monitoring in Europe (EAGLE). At the same time, globally performed exercises and fora are initiated from the C/GLC producers themselves and in coordination with international remote sensing associations in order to locally and regionally validate the land cover products. The latter occurs as a necessity to account for the products' dependence on a huge variety of geographical and climatic conditions and enhance credibility towards policy makers, stakeholders, and entrepreneurs. Confidence needs to be built up.

C/GLC maps represent the most important sources of accumulative and homogenized information about the surface of the earth and are used for several policies and scientific applications such as environmental monitoring, water monitoring, biodiversity, urban planning, and change detection of global land cover [3–5]. There are several C/GLC maps—such as IGBP-DISCover, GlobCover maps, MODIS GLC, LC-CCI maps and FROM-GLC maps [4]. Currently, many organizations produce C/GLC maps with higher resolution, namely the Land Cover-CCI (LC-CCI) maps at 300 m, GIO High Resolution Layers at 20 m, and Globeland at 30 m [4,6,7]. They all have been produced by remote sensing analysis using various optical data and methods. However, they are produced as independent datasets with different class hierarchies, semantic class similarities, and considerable disagreement among them have been reported as a consequence [3,4,8].

Comparative accuracy assessment of C/GLC maps, either one against another or juxtaposed against very high resolution ground data, is crucial but challenging, because of the lack of reference data. Several studies have assessed C/GLC products to analyze their weaknesses and strengths [3,9–11]. A few studies compared the accuracy estimates by harmonizing confusion matrices, but it remains unclear how they compared the C/GLC maps with the same reference dataset [3,4,8]. Reference datasets that are suitable for multiple maps were developed and used for validation of C/GLC maps [12]. However, these studies provide spatial agreement between C/GLC maps, but they neither compare the very recent high resolution products nor they estimate confidence levels during validation.

Triggered by the aforementioned challenges and recent developments, this study presents a quantitative and qualitative evaluation and inter-comparison for the recently produced C/GLCs, CORINE Land Cover 2012, GIO High Resolution Layers and Globeland30. The focus was on a representative landscape of the Northern Mediterranean basin, the area of Thessaly in Greece. The confidence levels of the experts were incorporated during the validation through a weighted overall accuracy assessment using manually annotated reference data, formed based on existing Google Earth images. In addition, the type of reported errors among the semantic classes is discussed, revealing further qualitative aspects of the considered C/GLCs.

2. Materials and Methods

2.1. Study Area

The study area is centered at 39°24′8.06″ N latitude and 21°59′1.10″ E longitude (Figure 1). It is located in central Greece near the regions of Macedonia, Epirus, Central Greece and borders with the Aegean Sea on the east. It covers an area of 14,036 square km (Figure 1), and includes (since *Kallikratis* reform of 2010) five second-level Local Administrative Units (LAU) (former NUTS 3-Nomenclature of Territorial Units for Statistics) and 25 municipalities. It also includes the Sporades Islands. Thessaly can be considered an optimal region for validating land cover mapping results, since it presents high landscape and land cover diversity, including islands and main land in the same area, mountains and plain areas and mixed land/use conditions. At the same time, it is a representative case of a typical Northeast Mediterranean landscape in terms of land cover and climatic conditions. In particular, the overall topography of the area consists of mountainous and semi-mountainous areas on the perimeter and lowlands in the center. The whole area includes five mountain massifs, containing the spur of Pindus that peaks up to Mount Olympus, with an altitude of 2917 m above sea level. The area is additionally comprised of rural and forested areas, with the Thessaly plain to

be one of the most important agricultural areas of the country. The rest of the land cover is divided into urban areas and ranching land. Moreover, Thessaly contains certain protected Natura2000 areas (e.g., Lake Karla), several statutory and non-protected areas, as well as one UNESCO World Heritage Site, namely Meteora [13].

Figure 1. Thessaly, the study area, is located in central Greece and presents high landscape and land cover diversity.

2.2. Data Sets

In this study, three openly available Land Cover (LC) maps are employed, namely the Copernicus CORINE Land Cover 2012 (CLC2012) map [14], the Pan-European Copernicus GIO High Resolution Layers (HRLs) of 2012 [15] and the GlobeLand30 (GLC30) [6]. In particular, CLC2012 represents the fourth CORINE Land Cover inventory. For its production, dual coverage of satellite images (IRS P6 LISS III and RapidEye) was used. It consists of an inventory of land cover in 44 classes. Corine Land Cover 2012 products are available in both raster (100 m and 250 m resolution) and vector (ESRI and SQLite geodatabase) formats in the European projection system (EPSG: 3035). Pan-European High Resolution Layers (HRLs) are designed as complementary to land cover/land use mapping, such as in the CORINE land cover (CLC) datasets [16]. The HRLs are produced from 20 m resolution satellite imagery through a combination of automatic processing and interactive rule based classification. The data is available as georeferenced raster data in the European projection (EPSG: 3035) with 20 m and 100 m spatial resolution. HRLs consist of five thematic classes, i.e., *Imperviousness*, *Forests*, *Grassland*, *Wetlands*, and *Permanent Water Bodies*, corresponding to main thematic classes of the CLC. GLC30 dataset production involved multispectral images with 30 m spatial resolution, including the TM5 and ETM+ of America Land Resources Satellite (Landsat) and the multispectral images of China Environmental Disaster Alleviation Satellite (HJ-1). The classification system includes 10 land cover types and uses as a baseline the year 2010 [6]. The data is available as georeferenced raster data in WGS 84 coordinate system, UTM projection, six-degree zoning and reference ellipsoid WGS 84 ellipsoid, with 30 m resolution [17].

During the validation process, the main objective was the accuracy assessment and inter-comparison of four LC categories, namely the (i) *Artificial Surfaces*; (ii) *Forest*; (iii) *Water*; and (iv) *Agriculture*. These four categories consist of LC classes that in most cases can be clearly distinguished from the interpreter and they are also available as separate classes on the three studied C/GLC (apart from category *Agriculture* for HRLs). These classes are juxtaposed against the 1. *Artificial Surfaces*, 3. *Forest*

and Semi-Natural Areas (including only the Level-2 class 3.1 *Forests*), 5. *Water Bodies* (including only the Level-2 subclass of 5.1 *Inland Waters*) and 2. *Agricultural Areas* from the CLC2012 Level-1 classes dataset. Regarding the HRLs dataset, the 20 m layers of *Imperviousness*, *Forest Type*, and *Permanent Water Bodies* were only employed, since HRLs do not offer a specific layer for agriculture. Regarding GLC30, the classes *Artificial Surfaces*, *Forest*, *Water Bodies*, and *Cultivated Land* were studied. To this end, Table 1 indicates the class/name correspondences among the different selected LC classes.

Table 1. The corresponding LC classes between the different employed LC maps.

General Categories	GLC30 2010-30 m	HRLs 2012-20 m	CLC2012 Level 1-100 m
Artificial Surfaces	Artificial Surfaces (code 80)	Imperviousness	1. Artificial Surfaces
Forest	Forest (code 20)	Forest Type	3. Forest and Semi-Natural Areas (only 3.1.)
Water	Water Bodies (code 60)	Permanent Water Bodies	5. Water Bodies (only 5.1.)
Agriculture	Cultivated Land (code 10)		2. Agricultural Areas

2.3. Accuracy Assessment and Validation Methodology

2.3.1. Sampling Design Based on CLC2012 Level-3

The accuracy assessment of the three different C/GLC maps was performed based on the inter-comparison with reference samples, which were derived after a comprehensive sampling procedure based on the CLC2012 product. In particular, random sampling points were generated automatically for each CLC2012 Level-3 (L3) subclass on the vector files under a stratified sampling design. These sampling points (X, Y) were then assigned to the nearest pixel (i.e., nearest pixel's center) for HRLs and GLC30 land cover product (see Section 2.3.3). CLC2012 L3 consists of subclasses with significantly higher level of detail than the ones of the other two products, so it was chosen for the sampling design basic scheme so as the stratification would ensure a precise estimate for each subclass.

The sampling unit was chosen to be one pixel for each land cover product since samples larger than the minimum mapping unit are frequently correlated with complicated cases of mixed samples, including more than one land-cover type [18]. Regarding the decision over the size of the sample that forms the testing set, the basic sampling theory was adopted to derive an estimate of the required sample size. More analytically, the approach proposed for defining the testing set in remote sensing studies was applied [19], using

$$n = \frac{z_{a/2}^2 \, P(1-P)}{h^2} \quad (1)$$

where n is the sample size, $z_{a/2}$ the critical value of the normal distribution for the two-tailed significance level α, P is a planning value of the population proportion for the correctly allocated cases, and h the half width of the desired confidence interval. Here, a typically adopted 0.05 significance level was considered, giving a $z_{a/2}$ equal to 1.96. A large conservative value for P of 0.5 was used and the target confidence interval was set between $\pm$4% up to $\pm$5%. For an h of 0.04 applying the Equation (1) an estimation of sample size of 601 samples is derived and for the value 0.05, an estimation of 385.

A stratified sampling scheme [4,19–22] was employed based on the area proportion of each L3 subclass to the total cover area, while applying constraints over maximum and minimum sizes for the subclasses. In order to remain within the required n range of 385–601 samples, a maximum of 120 and a minimum of 5 per L3 class was set. More analytically, CLC2012 class covering the highest relative extent for CLC2012 in the study area—i.e., 2.1.2. *Permanently irrigated land* with 25.03%—is given the maximum of 120 samples. All other Level-3 categories are attributed a sample number proportional to the above relationship and their relative extent. However, the minimum of five samples is given to each CLC 2012 L3 category, that was attributed with less than five samples, resulting to an overall sample size of 539 samples (inside the required range). Table 2 presents the percentage of the relative extent of each subclass and the derived number of samples, based on the procedure described above.

Table 2. The number of samples that were randomly collected per L3 CLC2012 subclass based on a stratified sampling methodology. Overall, more than 500 samples were collected, i.e., 62 for *Artificial Surfaces*, 338 for *Agriculture*, 129 for *Forest*, and 10 for the *Water* class.

CLC2012-Based Stratified Sampling				
CLC L1 Classes	**L3 Sub-Classes**	**Area Coverage**	**# of Samples Per Coverage**	**Selected # of Samples**
Artificial Surfaces	1.1.1. Continuous urban fabric	0.02%	0.07	5
	1.1.2. Discontinuous urban fabric	2.53%	12.12	12
	1.2.1. Industrial or commercial units	0.50%	2.41	5
	1.2.2. Road and rail networks and associated land	0.19%	0.89	5
	1.2.3. Port areas	0.00%	0.02	5
	1.2.4. Airports	0.26%	1.25	5
	1.3.1. Mineral extraction sites	0.12%	0.55	5
	1.3.2. Dump sites	0.00%	0.01	5
	1.3.3. Construction sites	0.03%	0.15	5
	1.4.1. Green urban areas	0.00%	0.02	5
	1.4.2. Sport and leisure facilities	0.03%	0.16	5
	Sum Artificial Surfaces:			***62***
Agriculture	2.1.1. Non-irrigated arable land	24.00%	115.02	115
	2.1.2. Permanently irrigated land	25.03%	120.00	120
	2.1.3. Rice fields	0.01%	0.07	5
	2.2.1. Vineyards	0.21%	1.03	5
	2.2.2. Fruit trees and berry plantations	0.80%	3.84	5
	2.2.3. Olive groves	2.62%	12.58	13
	2.3.1. Pastures	1.96%	9.40	9
	2.4.2. Complex cultivation patterns	4.60%	22.04	22
	2.4.3. Land principally occupied by agriculture. with significant areas of natural vegetation	9.11%	43.68	44
	Sum Agriculture:			***338***
Forest	3.1.1. Broad-leaved forest	13.76%	65.94	66
	3.1.2. Coniferous forest	7.94%	38.05	38
	3.1.3. Mixed forest	5.17%	24.77	25
	Sum Forest:			***129***
Water	5.1.1. Water courses	0.30%	1.45	5
	5.1.2. Water bodies	0.80%	3.82	5
	Sum Water:			***10***
	Total Sum:	***100.00%***	***Total Sum:***	***539***

2.3.2. HRLs and GlobeLand30 Data Pre-Processing

Standard data preprocessing techniques were required for a more efficient data/map management and direct comparison. Both maps were re-projected into the same coordinate reference system, i.e., the system of the two Copernicus datasets; ETRS89/ETRS-LAEA. Basic image algebra techniques were applied for multiplying the given HRLs layers with different constant numbers so as to label each LC class with a unique value after merging all the layers. The main goal here was to create a single raster file/map for each one of the HRLs and GLC30 datasets. Regarding the GLC30 dataset, the three raster files that contained the area of interest with all the classes and subclasses were mosaicked, and then the output file was re-projected from WGS 84/UTM zone 34/35N to ETRS89/ETRS-LAEA. It should be noted that a geolocation (mostly translation) disagreement of about 0.5–1 pixel was observed between the three different GLC30 raster files. At the same time, by optically comparing the given products at the coastline areas, one GLC30 raster file was found in better geolocation agreement with the Copernicus datasets (CLC2012, HRLs) than the other two raster files; thus, the GLC30 TIF files disagreement was handled during the merging procedure by shifting the two remaining raster files based on the first one. All in all, the final merged products for the three C/GLC datasets presented a mis-registration disagreement of about 0–3 HRL 20 m pixels, i.e., 0 m–60 m, in comparison with the coastline delineated with the images present in Google Earth. This is a fact that cannot be further improved neither by the

C/GLC layers user nor by the authorship team, as it would require involvement in the C/GLC layers original generation, which exceeds the objective of this study, i.e., to inter-compare the C/GLC layers following their production and as offered to the user.

2.3.3. Automated Class Label Retrieval

The sampling points that were randomly selected on the CLC2012 were used to retrieve automatically the corresponding values (labels) from the HRLs and GLC30 raster files. For this purpose, a MATLAB script was developed, which (i) reads the table with the geographic location of the sampling points (X, Y); (ii) reads the processed image files of HRLs and GLC30; (iii) stores the value (label) of the pixel whose center is nearest to the given X, Y; and (iv) creates a table with the sample id, the X, Y coordinates, and the corresponding labels from the HRLs and GLC30 layers. These files were further edited in order to contain also the LC labels in text.

2.3.4. Reference Data Annotation Based on Google Earth Images

The reference data were manually created after an intensive manual image interpretation procedure. For every sample two image interpretation experts assigned a LC label, i.e., *Artificial Surfaces*, *Forest*, *Water*, *Agriculture*, based on Google Earth very high resolution imaging data for the baseline years 2010 (GLC30) and 2012 (CLC2012 and HRLs) and a date as close as possible to the date of raw image acquisition for the generation of the C/GLCs. Especially for the *Water* class, the former prerequisite was given special attention, and sightings both prior and after the raw image acquisition date are additionally taken into account, using also the USGS Landsat Global Archive imagery when Google Earth Images were not available for these dates.

In particular, the procedure included the following steps: (i) the sampling points for each subclass (Level-3) of the CLC2012 were translated into kml files and then (ii) were inserted into Google Earth, so (iii) each expert assigned a LC label by interpreting the high-resolution Google Earth images of the same period. Since reference imagery of higher resolution is employed than the resolution of the maps assessed, the use of a smaller minimum mapping unit (MMU) is possible for the reference labeling [23]. The aim here was to interpret directly on the sampling point over Google Earth imagery for the reference classification. However, for cases that one sample was on/near the borders of two or more different classes the experts examined the surrounding area and dominating classes, using a patch of about 60 m $\times$ 60 m around the point in order to also account for the geolocation error observed between the different products (see Section 2.3.2).

At the same time, the experts recorded a confidence level indicating their scoring confidence for each and every annotation while generating the reference data base. Three confidence levels were assigned, i.e., #1 for >75% confidence level, #2 for 25–75% confidence level and #3 for confidence level below 25%. For all samples (approximately 250–300) that were annotated with a different label by the experts or were annotated with lower confidence levels (<75%), a second revision was performed by both experts in order to reach and agree to a common consensus.

2.3.5. Accuracy Assessment and Weighted Metrics

The performance of each LC map was validated forming confusion matrices against the manually annotated reference data. Agreement between the maps and reference classifications was defined by the class correspondence shown in Table 1. In order to exploit the level of confidence assigned by the experts' interpretation different confusion matrices were computed per confidence level. The standard accuracy metrics of Overall Accuracy (OA), User's and Producer's Accuracy (UA and PA), and kappa coefficient were calculated for each case. The incorporation of the confidence parameter to the validation process was achieved through Equation (2), which combines the accuracy metric

result to the respective weight of the level of confidence and number of observations in order to obtain a weighted overall accuracy by

$$wA = \frac{\sum_{i=1}^{3} w_i \, N_i \, A_i}{\sum_{i=1}^{3} w_i \, N_i} \quad (2)$$

where w_i is the given weight for the i confidence level, N_i the number of observations and A_i the achieved accuracy metric rate for the given confidence group.

The different weights are calculated using the median of the confidence level, defined based on the percentage range, i.e., 87.50 for 75–100% (#1), 50.00 for 25–75% (#2), and 12.50 for 0–25% (#3). Then the corresponding weight w_i for each confidence class is calculated weighting on the median, as 0.583 for confidence level #1, 0.333 for #2 and 0.083 for #3.

2.3.6. Inter-Comparison between the C/GLC Products

The quantitative and qualitative evaluation of the studied products was supplemented by the inter-comparison between the studied C/GLC layers. The inter-comparison was conducted through an assessment of the agreement between the three C/GLC products for each studied LC class. In particular, the inter-comparison assessment was achieved using simple image algebra, by producing difference images, after resampling all raster files to 20 m pixel size. The inter-comparison of all products required performing three groups of calculations for the particular differences: CLC2012-GLC30, CLC2012-HRLs, and CLC2012-HRLs, regarding all four classes in the first case and three classes (without *Agriculture*) for the last two comparisons. The agreement was assessed by calculating the fractions of the overlapping areas on both products for the corresponding LC classes (see Table 1).

3. Experimental Results and Validation

In this Section results from the validation of the three LC datasets against the manually annotated reference data are presented. In particular, in Section 3.1 results obtained from the accuracy assessment and weighted accuracy metrics calculation for all three products are presented. Then, in Section 3.2, the contribution of the confidence level that the experts assigned for each interpretation is analyzed. Lastly, an inter-comparison between the three C/GLC products is assessed in Section 3.3 by estimating the overlapping fraction of corresponding classes by pairs of products.

3.1. Validation Against the Reference Data

The three LC datasets were validated against the reference data through a quantitative accuracy assessment procedure. The confusion matrices for all three datasets against the manually annotated reference data of all confidence levels are presented on Tables 3–5. As it can be observed in all products' error matrices, OA rates higher than 89% were recorded for the majority of samples, which were annotated with confidence level #1, while OA for samples of lower confidence levels, exceeded 71%.

Table 3. The confusion matrix for CLC2012 validation against the reference data (RD) of classes: *Artificial Surfaces* (AS), *Agriculture* (AG), *Forest* (F), and *Water* (W), for samples of all confidence levels.

		CLC2012 Validation-Confidence Level: #1 | #2 | #3						
		AS	**AG**	**F**	**W**	**O/UN**	***Sum***	**PA (%)**
RD	**AS**	**33 | 9 | 2**	4 | 4 | 2	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	37 | 13 | 4	89 | 69 | 50
	AG	1 | 9 | 0	**164 | 129 | 10**	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	165 | 138 | 10	99 | 93 | 100
	F	0 | 14 | 0	0 | 0 | 1	**70 | 35 | 6**	0 | 0 | 0	0 | 0 | 0	70 | 49 | 7	100 | 71 | 86
	W	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	**7 | 1 | 0**	0 | 0 | 0	7 | 1 | 0	100 | 100 | -
	O/UN	2 | 5 | 1	3 | 7 | 0	3 | 12 | 3	2 | 0 | 0	**0 | 0 | 0**	10 | 24 | 4	0 | 0 | 0
	Sum	36 | 37 | 3	171 | 140 | 13	73 | 47 | 9	9 | 1 | 0	0 | 0 | 0	**289 | 225 | 25**	
	UA (%)	92 | 24 | 67	96 | 92 | 77	96 | 74 | 67	78 | 100 | -	-	**OA:**	**95% | 77% | 72%**
							kappa:	**0.91 | 0.60 | 0.58**

Table 4. The confusion matrix for HRLs validation against the reference data (RD) of classes: *Artificial Surfaces* (AS), *Forest* (F), and *Water* (W), for samples of all confidence levels.

		HRLs Validation-Confidence Level: #1 | #2 | #3					
		AS	**F**	**W**	**O/UN**	***Sum***	**PA (%)**
RD	**AS**	21 | 6 | 2	3 | 0 | 0	0 | 0 | 0	13 | 7 | 2	37 | 13 | 4	57 | 46 | 50
	F	0 | 0 | 0	70 | 41 | 5	0 | 0 | 0	0 | 8 | 2	70 | 49 | 7	100 | 84 | 71
	W	0 | 0 | 0	0 | 0 | 0	5 | 1 | 0	2 | 0 | 0	7 | 1 | 0	71 | 100 | -
	O/UN	0 | 2 | 0	7 | 12 | 2	0 | 0 | 0	168 | 148 | 12	175 | 162 | 14	96 | 91 | 86
	Sum	21 | 8 | 2	80 | 53 | 7	5 | 1 | 0	183 | 163 | 16	289 | 225 | 25	
	UA (%)	100 | 75 | 100	88 | 77 | 71	100 | 100 | -	92 | 91 | 75	**OA:**	**91% | 87% | 76%**
						kappa:	**0.84 | 0.70 | 0.56**

Table 5. The confusion matrix for GLC30 validation against the reference data (RD) of classes: *Artificial Surfaces* (AS), *Agriculture* (AG), *Forest* (F), and *Water* (W), for samples of all confidence levels.

		GLC30 Validation-Confidence Level: #1 | #2 | #3						
		AS	**AG**	**F**	**W**	**O/UN**	***Sum***	**PA (%)**
RD	**AS**	27 | 10 | 3	6 | 3 | 1	0 | 1 | 0	0 | 0 | 0	3 | 0 | 0	36 | 14 | 4	75 | 71 | 75
	AG	6 | 4 | 0	158 | 132 | 10	0 | 0 | 0	0 | 0 | 0	2 | 1 | 0	166 | 137 | 10	95 | 96 | 100
	F	0 | 0 | 0	0 | 13 | 1	73 | 25 | 10	0 | 0 | 0	0 | 3 | 0	73 | 41 | 11	100 | 61 | 91
	W	0 | 0 | 0	5 | 1 | 0	0 | 0 | 0	2 | 0 | 0	0 | 0 | 0	7 | 1 | 0	29 | 0 | -
	O/UN	3 | 0 | 0	4 | 7 | 1	1 | 15 | 4	0 | 0 | 0	1 | 3 | 0	9 | 25 | 5	11 | 12 | 0
	Sum	36 | 14 | 3	173 | 156 | 13	74 | 41 | 14	2 | 0 | 0	6 | 7 | 0	291 | 218 | 30	
	UA (%)	75 | 71 | 100	91 | 85 | 77	99 | 61 | 71	100 | - | -	17 | 43 | -	**OA:**	**90% | 78% | 77%**
							kappa:	**0.82 | 0.57 | 0.65**

By integrating all confidence levels together based on the proposed methodology (see Section 2.3.5), the resulting weighted OA reached the 89% for the CLC2012, the 90% for the HRLs and the 86% for the GLC30, while the weighted kappa coefficient estimation was 0.81, 0.79, and 0.74, respectively. Additional aspects regarding specific LC classes performance are derived from the weighted UA and PA rates of each class, presented in Figure 2. In particular, one can observe that in all cases the *Agriculture* and *Forest* classes had PA and UA of above 85%, indicating high accuracy and reliability, respectively.

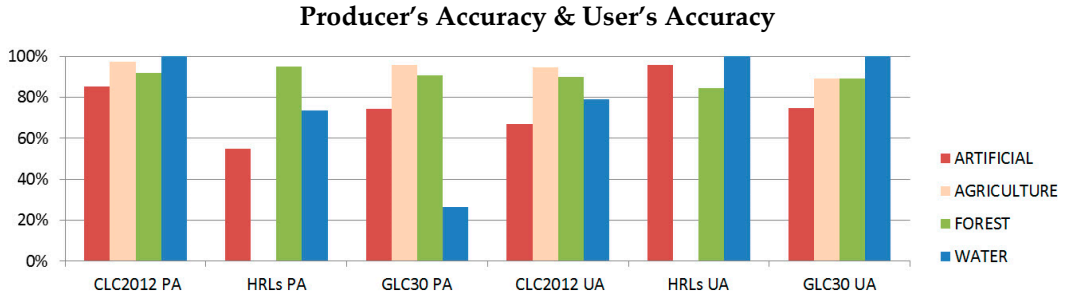

Figure 2. Weighted PA (Producer's Accuracy) and UA (User's Accuracy) for all studied LC datasets.

For CLC2012 the PA for *Artificial Surfaces* was 85% and 100% for *Water*; however, lower rates were recorded for the UA, i.e., 67% and 79%, respectively. GLC30 presented relatively high accuracy for the class *Artificial Surfaces* both for PA (74%) and UA (75%) metrics. However, GLC30 presented mis-classification cases for the class *Water*; the majority of reference data samples (six out of eight) were classified as *Agriculture* (Table 5), resulting in a weighted PA of 26%. For HRLs, several reference data samples of *Artificial Surfaces* were attributed to class *Others/Unclassified* reaching a relatively low PA rate of 55%. The class *Water* resulted into a 100% UA rate and a PA rate of 74% for the validation of HRLs.

In Table 6, the highest weighed scores recorded for PA and UA rates are presented per LC class. As it can be observed, in five out of eight cases the CLC2012 product validation achieved the highest values.

Table 6. The highest weighted PA (Producer's Accuracy) and UA (User's Accuracy) rates per LC class.

LC Class	PA	UA
Artificial Surfaces	ClC2012 (85%)	HRLs (96%)
Agriculture	ClC2012 (97%)	ClC2012 (95%)
Forest	HRLs (95%)	ClC2012 (90%)
Water	ClC2012 (100%)	HRLs & GLC30 (100%)

3.2. The Contribution of the Confidence Indicator

In this Section, the OA (and weighted OA) rates are compared both without and by taking into account the confidence level that the expert assigned at every reference data point during the image interpretation. In Table 7, the OA rates per confidence level for all LC maps are presented along with the number of samples per confidence level. The last two rows present the OA and the weighted OA rates. The latter are increased by 1–3% compared to the OA ones. In particular, as the confidence level decreases the OA rate decreases, too. This is quite expected, since the annotations with lower confidence levels, indicating difficulty in the labelling decision, usually involve particular regions, terrain types, and complex land cover/use cases, lying most probably on the borders of two LC classes, and therefore they are more associated with classification errors.

Table 7. Overall Accuracy and weighted Overall Accuracy rates per product.

LC Map	**ClC2012**			**HRLs**			**GLC30**		
Confidence Level	*CL1*	*CL2*	*CL3*	*CL1*	*CL2*	*CL3*	*CL1*	*CL2*	*CL3*
# of samples	*289*	*225*	*25*	*289*	*225*	*25*	*291*	*218*	*30*
OA per CL	*95%*	*77%*	*72%*	*91%*	*87%*	*76%*	*90%*	*78%*	*77%*
OA		**86%**			**89%**			**84%**	
Weighted OA		**89%**			**90%**			**86%**	

Similar conclusions are derived when comparing the resulting PA and UA per class rates for all products validation. In Table 8 one can also observe the differences between the standard and weighted PA and UA. For most cases the weighted PA and UA are increased around 1–4%. Classes *Water* and *Agriculture* present the smallest differences. A greater difference occurs for the UA of CLC2012 product validation of *Artificial Surfaces*, which presents a UA rate of 58% and a weighted UA rate of 67%. This 9% difference occurs due to the fact that a rather large number of *Agriculture* and *Forest* samples in the reference data sets (Table 3), characterized with a confidence level of 2, were annotated as *Artificial Surfaces* in the CLC2012 map. The contribution of these errors decreased when the weighted UA metric was calculated, since these confidence level #2 observations are given a decreased weight in the calculation.

Table 8. The calculated difference (%) between the standard PA and UA and weighted PA and UA (wPA, wUA) rates.

LC Maps	**ClC2012**		**HRLs**		**GLC30**	
	wPA-PA	**wUA-UA**	**wPA-PA**	**wUA-UA**	**wPA-PA**	**wUA-UA**
Artificial Surfaces	4%	9%	1%	2%	0%	−1%
Agriculture	1%	1%			0%	1%
Forest	4%	4%	3%	2%	4%	6%
Water	0%	−1%	−1%	0%	1%	0%

3.3. Inter-Comparison between the C/GLC Products

Apart from the comparison with the truth (reference data), useful aspects of the studied products derive from the inter-comparison between one another. To this end, an assessment of the agreement between the three C/GLC products was employed by comparing the overlapping fractions for the studied classes.

In Figure 3 the resulting images for the three mathematical differences, i.e, CLC2012-GLC30, CLC2012-HRLs, and GLC30-HRLs are presented, along with the proportional fraction of pixels attributed in the examined class: (i) on both products, (ii) only on the first product and (iii) only on the second product.

The comparison between CLC2012 and GLC30 reveals a high percentage of agreement between the two products for the *Agriculture* class (91%) and a quite high rate of 67% for the *Forest* class.

These specific classes scored also high PA and UA rates of above 85%, as analyzed in the previous paragraphs. Still 28% of all *Forest* pixels, are attributed to *Forest* class only for GLC30. These areas, located on the northern part of the study area are annotated as classes of the Level-2 *Scrub and/or Herbaceous Vegetation Associations* category in CLC2012. Similarly, 29% of all *Artificial Surfaces* pixels are attributed to this class only for GLC30, while both products share a common *Artificial Surfaces* area of 54%. The comparison for *Water* class recorded only a 26% shared fraction. This can be also linked with the low PA rates for this class in the GLC30 (see Section 3.1), since as one can observe in Figure 3, this product annotates Lake Karla as *Cultivated Land*

The comparison between CLC2012 and HRLs presents high discrepancy for class *Artificial Surfaces* (only 38% shared fraction). Many disagreement cases are mainly observed as a result of the scale difference of the two products, but can be also attributed to semantic differences between the corresponding classes and subclasses of *Imperviousness* (HRLs) and *Artificial Surfaces* (CLC2012). Furthermore, regarding the *Forest* class, 43% of all *Forest* pixels are characterized as *Forest* only for the HRLs product. These dissimilarity cases are mainly located in areas where CLC2012 subclasses of the Level-2 category *Scrub and/or Herbaceous Vegetation Associations* can be found. For *Water* class a shared fraction of 57% is recorded between the two products, since as it can be observed in the map, the northwestern part of Lake Karla has not been attributed to *Water* class in the HRLs product. Respectively, only a small percentage (7%) is characterized as *Water* only on the HRLs product and can be attributed to the detection of narrow linear parts of courses that may not be recorded on the coarser 100 m product of CLC2012.

The difference images for the comparison between GLC30 and HRLs also present high disagreement for class *Artificial Surfaces* (only 36% shared fraction), which can be also attributed to the semantic and classification methodology diversities between the corresponding classes of the two products. As in the previous comparison regarding the *Forest* class, 31% of all pixels are characterized as *Forest* only for the HRLs product. Areas of differences are mainly located in regions characterized in the GLC30 product as *Cultivated Land* and *Shrubland*. At last, disagreement recorded for *Water* class (only 37% shared fraction) are associated with the omission of Lake Karla from the GLC30 product.

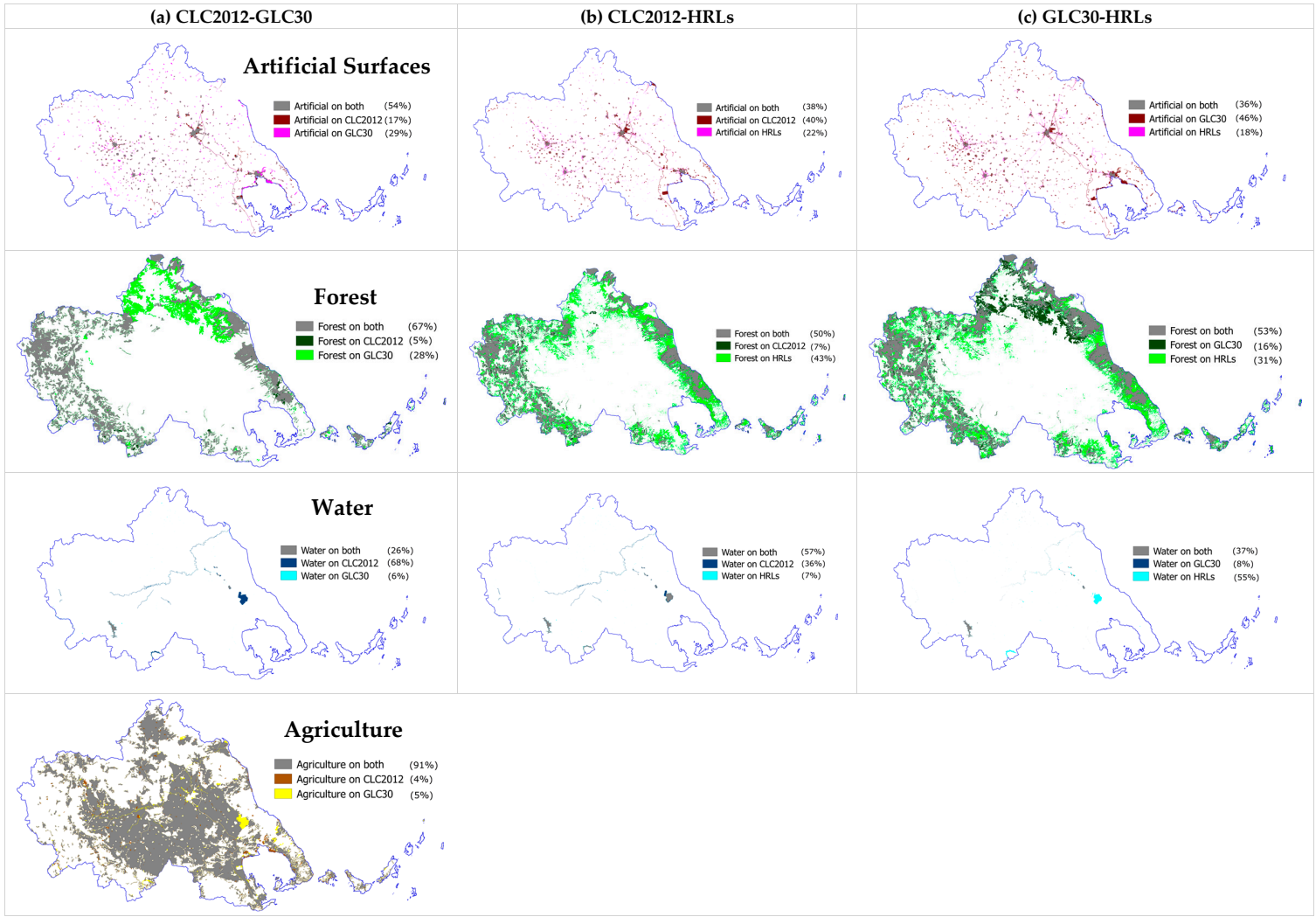

Figure 3. Qualitatively comparing the studied products by calculating the per-class differences between: (**a**) CLC2012-GLC30, (**b**) CLC2012-HRLs, and (**c**) GLC30-HRLs.

4. Discussion

Although various validation frameworks for the qualitative and quantitative assessment of C/GLCs have been presented in several studies [24–29] they neither used a common reference layer for the comparison nor did they incorporated a confidence level during the reference data production. Regarding the sampling design strategy, there are several arguments promoting for sample size increment as much as possible in relation with the reference layer. The spatial extent of an area is the crucial determining factor, as its relation with a higher spectral variability, especially in view of the need of image mosaicking required and the bidirectional reflectance distribution effects. Research studies arguing over the optimal sampling design and size determination [19,23] have guided the presented work to undertake a cost-effective stratified sampling approach. Thus, the four Level-1 CLC2012 classes were downscaled to their 25 Level-3 constituents, present in the study area, to account for the spectral variability of the four prevailing classes. This way, the percentiles of the extent of each of the 25 subclasses defined the variation and number of samples assigned to the four Level-1 classes, approaching the existing spectral variability through the appearance of the subclasses.

Moreover, the incorporation of the confidence level of the image interpreter within this process took into account the boundaries' effects in a weighted and non-negligible way; thus, being closer to reality. Using solely the certain location points would have biased the result towards a non-fully objective judgement, producing a higher performative OA result for the three studied layers, e.g., 95, 91, and 90% instead of the 89, 90, and 86% (see Tables 3–5 and 7). These reported quantitative results are similar with the ones presented in [25,26] and disagree with the relative lower OA rates (<50%) presented in [29]. The integration of the expert's confidence into the reference data annotation procedure has been already proposed and highlighted as a good practice for accuracy assessment in the literature [4,23,30]. Low, moderate and high confidence rates were used in the analysis in order to subset the results by confidence [18,31]. The visual interpretation here has been proven to be highly biased depending on the interpreter, as also stated in similar studies [4,30], so for every sample interpreted with lower confidence level or for cases that the two experts disagreed on labelling, a second round of interpretation took place in order to reach consensus. It should be also noted that most of the automatically and randomly selected samples were annotated with high or medium confidence levels and only a 5% of the samples was annotated with the lower confidence level #3. Since similar studies have indicated that OA results from a confusion matrix should be interpreted with caution, as the matrix records the degree of agreement between the reference data and the map data, which are in cases less than perfect [18,32], it is suggested here that the integration of confidence levels during annotation can effectively address these concerns.

Producers' and users' accuracy rates from the weighted accuracy assessment along with the direct inter-comparison of the products per pairs indicated that classes *Agriculture* and *Forest* scored high in accuracy (PA) and reliability (UA) while also presenting the highest overlapping fractions between the products. Although class *Water* presented high level of reliability for all products, omission errors on the GLC30 product for the water body of Lake Karla were reported and further highlighted from the difference images of the inter-comparisons. It should be noted that Lake Karla was listed among the swallow lakes in Greece up to 1962 when it was completely drained to gain land for agriculture [33]. Lake Karla's restoration project was launched in 2000 while its refilling started in 2009 [34]. Google Earth and Landsat imagery close to the raw image acquisition for GLC30 product (July 2009) picture the lake covered with water in the largest percentage of its current extent. The weighted accuracy assessment and inter-comparisons also indicated many commission errors and low overlapping fractions for class *Artificial Surfaces*. This particular class presents diversity on semantic definition of corresponding classes on the different products. In particular, HRLs include only sealed areas of imperviousness, while the other two products also account for open sites, e.g., mines and quarries and partially vegetated areas.

5. Conclusions

The validation procedure proposed in this study provides a qualitative and statistical outcome comparing the accuracy of three C/GLC products in detecting artificial, forest, agriculture, and (inland) water land cover classes for a study area in central Greece, Thessaly. The experts' confidence level per sample was recorded during reference data annotation and integrated in the evaluation process through a weighted accuracy assessment procedure. Validation results recorded high OA rates of 89, 90, and 86% for CORINE Land Cover 2012, GIO High Resolution Layers and Globeland30 datasets, respectively. Further analysis on the reported PA and UA identified certain classes, i.e., *Agriculture* and *Forest*, as the most accurate and reliable, possessing, moreover, the highest overlapping fractions during the inter-comparison. Lower accuracy rates were recorded for *Artificial Surfaces* most probably due to differences in spatial resolution and semantic definitions. The main conclusion of this work outlines that different quality aspects of the considered LC maps were noted and highlighted more transparently and objectively as a result of the employed confidence levels per sample, the stratified sampling and the weighted OA calculation.

Acknowledgments: We would like to thank the anonymous reviewers for the attentive review and constructive comments. We gratefully acknowledge support from the Research Projects for Excellence IKY (State Scholarships Foundation)/SIEMENS. Authors would like to acknowledge that the presented work is generated in the framework of the "Global Land Cover Products Validation and Inter-Comparison in South Central and Eastern Europe" activities within the South Central and Eastern European Regional Information Network (SCERIN).

Author Contributions: I.M. and K.K. conceived and designed the experiments; X.K. and I.G. performed the experiments; X.K. and I.G. analyzed the data; I.M. and K.K. contributed materials/analysis tools; I.M., X.K., K.K. and I.G. wrote and edited the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, J.; Mooney, H.; Hull, V.; Davis, S.J.; Gaskell, J.; Hertel, T.; Lubchenco, J.; Seto, K.C.; Gleick, P.; Kremen, C. Systems integration for global sustainability. *Science* **2015**, *347*, 1258832. [CrossRef] [PubMed]
2. Karantzalos, K.; Bliziotis, D.; Karmas, A. A scalable geospatial web service for near real-time, high-resolution land cover mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4665–4674. [CrossRef]
3. Fritz, S.; See, L.; McCallum, I.; Schill, C.; Obersteiner, M.; Van der Velde, M.; Boettcher, H.; Havlík, P.; Achard, F. Highlighting continued uncertainty in global land cover maps for the user community. *Environ. Res. Lett.* **2011**, *6*, 044005. [CrossRef]
4. Tsendbazar, N.; de Bruin, S.; Herold, M. Assessing global land cover reference datasets for different user communities. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 93–114. [CrossRef]
5. Verburg, P.H.; Neumann, K.; Nol, L. Challenges in using land use and land cover data for global change studies. *Glob. Chang. Biol.* **2011**, *17*, 974–989. [CrossRef]
6. Chen, J.; Chen, J.; Liao, A.; Cao, X.; Chen, L.; Chen, X.; He, C.; Han, G.; Peng, S.; Lu, M. Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 7–27. [CrossRef]
7. Mora, B.; Tsendbazar, N.-E.; Herold, M.; Arino, O. Global land cover mapping: Current status and future trends. In *Land Use and Land Cover Mapping in Europe*; Springer: Dordrecht, The Netherlands, 2014; pp. 11–30.
8. Herold, M.; Mayaux, P.; Woodcock, C.; Baccini, A.; Schmullius, C. Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. *Remote Sens. Environ.* **2008**, *112*, 2538–2556. [CrossRef]
9. Giri, C.; Zhu, Z.; Reed, B. A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets. *Remote Sens. Environ.* **2005**, *94*, 123–132. [CrossRef]
10. Manakos, I.; Chatzopoulos-Vouzoglanis, K.; Petrou, Z.I.; Filchev, L.; Apostolakis, A. Globalland30 mapping capacity of land surface water in Thessaly, Greece. *Land* **2014**, *4*, 1–18. [CrossRef]
11. McCallum, I.; Obersteiner, M.; Nilsson, S.; Shvidenko, A. A spatial comparison of four satellite derived 1 km global land cover datasets. *Int. J. Appl. Earth Obs. Geoinform.* **2006**, *8*, 246–255. [CrossRef]

12. Bontemps, S.; Defourny, P.; Bogaert, E.V.; Arino, O.; Kalogirou, V.; Perez, J.R. GLOBCOVER 2009-Products Description and Validation Report; UCLouvain and ESA 2011. Available online: http://due.esrin.esa.int/files/GLOBCOVER2009_Validation_Report_2.2.pdf (accessed on 30 November 2015).
13. Directorate of Spatial Planning. *Evaluation, Revision and Specialization of the Regional Framework for Spatial Planning and Sustainable Development of Thessaly Region*; Ministry of the Environment, Energy and Climate Change, General Secretariat of Regional Planning and Urban Development: Athens, Greece, 2013; pp. 1–122.
14. EEA—European Environment Agency. Corine Land Cover (CLC) 2012, Version 18/3. Available online: http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012/view (accessed on 30 November 2015).
15. EEA—European Environment Agency. Pan-European High Resolution Layers (HRL). Available online: http://land.copernicus.eu/pan-european/high-resolution-layers/view (accessed on 30 November 2015).
16. EEA—European Environment Agency. GIO land (GMES/Copernicus initial operations land) High Resolution Layers (HRLs)—Summary of product specifications, Version 7. Available online: http://land.copernicus.eu/user-corner/publications/gio-land-high-resolution-layers (accessed on 30 November 2015).
17. NGCC—National Geomatics Center of China. GlobeLand30. Available online: http://www.globallandcover.com/GLC30Download/download_t.aspx (accessed on 30 November 2015).
18. Powell, R.; Matzke, N.; De Souza, C.; Clark, M.; Numata, I.; Hess, L.; Roberts, D. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sens. Environ.* **2004**, *90*, 221–234. [CrossRef]
19. Foody, G.M. Sample size determination for image classification accuracy assessment and comparison. *Int. J. Remote Sens.* **2009**, *30*, 5273–5291. [CrossRef]
20. Foody, G.M.; Pal, M.; Rocchini, D.; Garzon-Lopez, C.X.; Bastin, L. The sensitivity of mapping methods to reference data quality: Training supervised image classifications with imperfect reference data. *ISPRS Int. J. Geo-Inform.* **2016**, *5*, 199. [CrossRef]
21. Mayaux, P.; Eva, H.; Gallego, J.; Strahler, A.H.; Herold, M.; Agrawal, S.; Naumov, S.; De Miranda, E.E.; di Bella, C.M.; Ordoyne, C. Validation of the global land cover 2000 map. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 1728–1739. [CrossRef]
22. Silva, J.; Bacao, F.; Foody, G.; Caetano, M. *Automatic Selection Of Training Areas Using Existing Land Cover Maps*; ESA Special Publication: Paris, France, 2013; p. 184.
23. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
24. Arsanjani, J.J.; Tayyebi, A.; Vaz, E. GlobeLand30 as an alternative fine-scale global land cover map: Challenges, possibilities, and implications for developing countries. *Habitat Int.* **2016**, *55*, 25–31. [CrossRef]
25. Brovelli, M.A.; Molinari, M.E.; Hussein, E.; Chen, J.; Li, R. The first comprehensive accuracy assessment of GlobeLand30 at a national level: Methodology and results. *Remote Sens.* **2015**, *7*, 4191–4212. [CrossRef]
26. Caetano, M.; Mata, F.; Freire, S.; Campagnolo, M. Accuracy assessment of the Portuguese CORINE Land Cover map. In *Global Developments in Environmental Earth Observation from Space*; Marcal, A., Ed.; Millpress: Rotterdam, The Netherlands, 2006; pp. 459–467.
27. Pérez-Hoyos, A.; García-Haro, F.; San-Miguel-Ayanz, J. Conventional and fuzzy comparisons of large scale land cover products: Application to CORINE, GLC2000, MODIS and GlobCover in Europe. *ISPRS J. Photogramm. Remote Sens.* **2012**, *74*, 185–201. [CrossRef]
28. Sannier, C.; Gallego, J.; Dahmer, J.; Smith, G.; Dufourmont, H.; Pennec, A. Validation of Copernicus high resolution layer on imperviousness degree for 2006, 2009 and 2012. In Proceedings of the International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Montpellier, France, 5–8 July 2016; pp. 5–8.
29. Sun, B.; Chen, X.; Zhou, Q. Uncertainty Assessment of GLOBELAND30 Land Cover Data Set Over Central Asia. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *XLI-B8*, 1313–1317.
30. Strahler, A.H.; Boschetti, L.; Foody, G.M.; Friedl, M.A.; Hansen, M.C.; Herold, M.; Mayaux, P.; Morisette, J.T.; Stehman, S.V.; Woodcock, C.E. *Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps*; European Communities: Luxembourg, 2006; Volume 51.
31. Wickham, J.; Stehman, S.; Fry, J.; Smith, J.; Homer, C. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. *Remote Sens. Environ.* **2010**, *114*, 1286–1296. [CrossRef]
32. Foody, G.M. Status of land cover classification accuracy assessment. *Remote sens. Environ.* **2002**, *80*, 185–201. [CrossRef]

33. Chamoglou, M.; Papadimitriou, T.; Kagalou, I. Key-descriptors for the functioning of a Mediterranean reservoir: The case of the New lake Karla-Greece. *Environ. Process.* **2014**, *1*, 127–135. [CrossRef]
34. Dodouras, S.; Lyratzaki, I.; Papayannis, T. *Lake Karla Walking Guide*; Med-INA: Athens, Greece, 2014.