Committee on Earth Observation Satellites

Working Group on Calibration and Validation

Land Product Validation Subgroup

Land Cover Focus Area

Land Cover and Change Map Accuracy Assessment and Area Estimation Good Practices Protocol

Version 1.1 - 2025

Editors: Alexandra Tyukavina, Stephen V. Stehman, Giles M. Foody, Sophie Bontemps, Anna Komarova, Nandin-Erdene Tsendbazar, Jaime E. Nickeson

Chapter leads: Alexandra Tyukavina (Chapters 1 - 5), Sophie Bontemps (Chapters 1, 2, Appendix), Pontus Olofsson (Chapters 3, 5), Giles M. Foody and Julien Radoux (Chapter 4), Linda See and Bryant M. Serre (Chapter 6), Xiao-Peng Song (Chapter 7)

Citation: Tyukavina, A., Stehman, S. V., Foody, G. M., Bontemps, S., See, L., Olofsson, P., Tsendbazar, N. E., Radoux, J., Komarova, A., Serre, B. M., Song, X. P., d'Andrimont, R., Koren, G., Potapov. P., Bullock, E. L., Campbell., P., de Bruin, S., Defourny, P., Friedl., M. A., Fritz., S., Hansen, M. C., Herold, M., Lamarche, C., Lesiv, M., Mané, L., Meroni, M., Nickeson, J. E., Pelletier, F., Pickens, A., Reiche, J., Schepaschenko, D., Tarrio, K., Verhegghen, A., Woodcock, C., Xiao, X. (2025). Land Cover and Change Map Accuracy Assessment and Area Estimation Good Practices Protocol. Version 1.1. In A. Tyukavina, S. V. Stehman, G. Foody, S. Bontemps, A. Komarova, N. E. Tsendbazar and J. Nickeson (Eds.), Good Practices for Satellite Derived Land Product Validation, (p. 187): Land Product Validation Subgroup (WGCV/CEOS), doi:10.5067/doc/ceoswgcv/lpv/lc.001

Authors: Alexandra Tyukavina[1], Stephen V. Stehman[2], Giles M. Foody[3], Sophie Bontemps[4], Linda See[5], Pontus Olofsson[6], Nandin-Erdene Tsendbazar[7], Julien Radoux[4], Anna Komarova[1], Bryant M. Serre[8], Xiao-Peng Song[1], Raphaël d'Andrimont[9], Gerbrand Koren[10], Peter Potapov[1], Eric L. Bullock[11], Petya Campbell[12,13], Sytze de Bruin[7], Pierre Defourny[4], Mark A. Friedl[14], Steffen Fritz[5], Matthew C. Hansen[1], Martin Herold[7,15], Céline Lamarche[4], Myroslava Lesiv[5], Landing Mané[16], Michele Meroni[9], Jaime E. Nickeson[12,17], Flavie Pelletier[8], Amy H. Pickens[1], Johannes Reiche[7], Dmitry Schepaschenko[5], Katelyn Tarrio[14], Astrid Verhegghen[9], Curtis E. Woodcock[14], Xiangming Xiao[18]

[1] Department of Geographical Sciences, University of Maryland, College Park, MD, USA
[2] College of Environmental Science and Forestry, State University of New York, Syracuse, NY, USA
[3] School of Geography, University of Nottingham, Nottingham, UK
[4] Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[5] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[6] NASA Marshall Space Flight Center, Huntsville, AL, USA
[7] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen, the Netherlands
[8] Department of Natural Resource Sciences, McGill University, Montreal, Canada
[9] European Commission, Joint Research Centre (JRC), Ispra, Italy
[10] Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, the Netherlands
[11] Rocky Mountain Research Station, U.S. Forest Service, Riverdate, UT, USA
[12] NASA Goddard Space Flight Center, Biospheric Sciences Laboratory, Greenbelt, MD, USA
[13] Goddard Earth Sciences Technology and Research (GESTAR) II, University of Maryland Baltimore County, Baltimore, MD, USA
[14] Department of Earth and Environment, Boston University, Boston, MA, USA
[15] Helmholtz GFZ German Research Centre for Geoscience, Remote Sensing and Geoinformatics Section, Telegrafenberg, Potsdam, Germany
[16] Observatoire Satellital des Forêts d'Afrique Centrale (OSFAC), Kinshasa, Democratic Republic of the Congo
[17] Science Systems and Applications, Inc., Lanham, MD, USA
[18] Department of Microbiology and Plant Biology, Center for Earth Observation and Modeling, University of Oklahoma, Norman, OK, USA

We are grateful to the expert reviewers whose thoughtful comments and constructive suggestions contributed to the improved version 1.0 of the document.

Reviewers: Frédéric Achard[1], Gilberto Camara[2], René Colditz[1], Nicholas Cuba[3], Qiongyu Huang[4], Fabrizio Niro[5], Leandro Leal Parente[6], George P. Petropoulos[7], Anna Pustogvar[8], Daniel Sousa[9], Le Yu[10], Viviana Zalles[11]

[1] European Commission, Joint Research Centre (JRC), Ispra, Italy
[2] National Institute for Space Research (INPE), Brazil
[3] Auburn University at Montgomery, Montgomery, AL, USA
[4] Conservation Biology Institute, Smithsonian Institution, Front Royal, VA, USA
[5] Serco for European Space Agency Centre for Earth Observation (ESA-ESRIN), Frascati, Italy
[6] OpenGeoHub Foundation, Wageningen, the Netherlands
[7] Department of Geography, Harokopio University of Athens, Athens, Greece
[8] National Physical Laboratory, University of Leicester, Leicester, UK
[9] San Diego State University, San Diego, CA, USA
[10] Department of Earth System Science, Tsinghua University, Beijing, China
[11] World Resources Institute, Washington D.C., USA

List of Revisions

Version	Revision	Date
0.1	Draft released for community review	August 30, 2024
1.0	Reviewers' comments addressed; revised version published on the CEOS WGCV LPV website	September 15, 2025
1.1	This version incorporates revisions from and is endorsed by CEOS WGCV	November 2025

Table of Contents

EXECUTIVE SUMMARY **6**

1. INTRODUCTION **8**

1.1 SCOPE OF THE GUIDELINES 8
1.2 THE CEOS LPV VALIDATION STAGES 9
1.3 CURRENT STATE OF GLOBAL AND CONTINENTAL-SCALE LAND COVER AND CHANGE MAPPING AND VALIDATION 12
1.4 GLOBAL LAND COVER MAPS 14
1.5 REQUIREMENTS FOR LAND COVER ESSENTIAL CLIMATE VARIABLE (ECV) 16

2. DEFINITIONS AND GENERAL PRINCIPLES **22**

2.1 DEFINITIONS 22
2.2 KEY PRINCIPLES OF ACCURACY ASSESSMENT 31
2.3 LAND COVER, LAND COVER CHANGE, LAND USE 36
2.4 CATEGORICAL MAPS VS. CONTINUOUS FIELDS 39
2.5 LAND COVER CHANGE MAPS 41
2.6 MAP ACCURACY ASSESSMENT VS. OTHER MAP AND MODEL QUALITY ASSESSMENT METHODS 43
2.7 ACCURACY METRICS AND AREA ESTIMATES 46
2.8 COMPARATIVE MAP ACCURACY ASSESSMENT 51
2.9 INTERCOMPARISON OF MAPS 53
2.10 SYSTEMATIC MAP QUALITY CONTROL 54

3. SAMPLING DESIGN **55**

3.1 SAMPLING UNIT 57
3.2 SAMPLING FRAME 59
3.3 COMMON PROBABILITY SAMPLING DESIGNS 60
3.4 STRATIFICATION 64
3.5 SAMPLE SIZE PLANNING AND ALLOCATION TO STRATA 66

4. RESPONSE DESIGN **75**

4.1 SAMPLE LABELING PROTOCOL 80
4.2 QUALITY OF REFERENCE DATA 83
4.3 ACCOUNTING FOR REFERENCE DATA ERRORS 86

5. ANALYSIS **91**

5.1 ESTIMATING MAP ACCURACY 92
5.2 ESTIMATING TARGET CLASS AREA 97

6. SOURCES OF REFERENCE DATA **97**

6.1 TIME-SERIES OF MEDIUM- TO VERY HIGH-RESOLUTION OPTICAL DATA 103
6.2 SPACEBORNE AND AIRBORNE LIDAR DATA 107
6.3 DATA FROM UAV 109
6.4 GROUND SURVEYS 110
6.5 EXPERT-BASED METHODS OF REFERENCE DATA COLLECTION VS. CROWDSOURCING 115

7. CHALLENGES AND FUTURE DIRECTIONS 119

7.1 OPERATIONAL VALIDATION UPDATES 120
7.2 ASSESSING ACCURACY OF NEAR REAL-TIME MAPS 123
7.3 TOWARD MORE STANDARDIZED VALIDATION DATASETS AND COLLECTIONS OF REFERENCE DATA 127
7.4 LOCAL MAP QUALITY METRICS 133

APPENDIX. EXAMPLES OF NATIONAL-, REGIONAL- AND GLOBAL-SCALE VALIDATION EFFORTS 136

A.1. VALIDATION OF THE 'GLOBAL LAND COVER 100M' FROM THE COPERNICUS GLOBAL LAND SERVICE 136
A.2. ESA CLIMATE CHANGE INITIATIVE GLOBAL LAND COVER TIME SERIES 141
A.3. ESA CLIMATE CHANGE INITIATIVE WATER BODIES PRODUCT 144
A.4. UMD GLAD VALIDATION OF SINGLE- AND MULTI-CLASS LAND COVER AND CHANGE MAPS 147
A.5. VALIDATION OF THE EUROPEAN CROP MAP 2018 151
A.6. VALIDATION ACTIVITIES WITHIN THE SATELLITE OBSERVATORY OF CENTRAL AFRICAN FORESTS (OSFAC) CONTEXT 155
A.7. VALIDATION STRATEGY FOR LAND COVER AND LAND COVER CHANGE IN SUPPORT OF GLANCE 158
A.8. VALIDATION OF THE CROPLAND MAPS 159

REFERENCES 163

Executive Summary

- The main purpose of these community good practice guidelines is to **encourage correct implementation** of land cover map accuracy assessment and area estimation methods by map producers and to enable map users to correctly interpret the provided map accuracy information. To do so, we provide an overview of key principles of accuracy assessment (section 2.2), recommend specific methods for various components of accuracy assessment published in high quality peer-reviewed studies, and point to potential misapplications of the presented methods that should be avoided.

- Advancements in remote sensing data acquisition, increased access to data and computational resources, and novel machine learning algorithms have resulted in an **increasing number of published land cover and change maps** of increasingly higher spatial resolution (Chapter 1). Since the early 2000s, improved access to satellite data archives and awareness in the land cover mapping community about validation good practices led to an increasing number of published maps being accompanied by accuracy assessments. However, **in many cases accuracy assessment methodology is not well described and reference data are not public** (Table 1.4), preventing independent verification of map validation quality.

- **All land cover and change maps contain errors, the magnitude of which needs to be assessed by comparing map labels with an independent reference sample**, which is referred to in this document as map **'accuracy assessment'** or **'validation'**. Map uncertainty that is quantified using algorithm variance metrics is not considered an accuracy assessment (section 2.6). A map that is lacking an accuracy assessment is just a prototype or an untested hypothesis and should not be used as a source of information.

- Land cover maps that have been validated only at the global scale should be used with caution for regional-, national- and local-scale assessments as **map accuracy varies in space**; map users are encouraged to perform a standard accuracy assessment described in the current document (section 2.7, Chapters 3-5) for their sub-region of interest to identify whether the global map meets the user-required accuracy within that sub-region.

- **Accuracy assessment should be designed to fit specific land cover map types** (e.g., categorical maps vs. continuous fields) **and assessment purposes** (e.g., standalone map validation vs. comparative accuracy assessment of multiple maps), see Chapter 2.

- **Accuracy assessment needs to be taken seriously,** with a significant budget (at least 30% of the total budget) allocated for this process when map production is being planned (section 2.2.1).
- **Recommended map accuracy assessment approach is based on a probability sample of reference data** (design-based inference framework). Chapters 3-5 outline the main accuracy assessment components: sampling design (Chapter 3), response design (Chapter 4) and analysis (Chapter 5).
- **The CEOS LPV validation stage 3 should be the goal for all newly published global land cover maps** (see section 1.2 for the description of the CEOS validation stages). Updating accuracy assessment for new map version releases and time-series expansion (validation stage 4) is recommended (section 7.1).
- **Reference data uncertainty needs to be quantified,** and either minimized or incorporated into the estimates (Chapter 4). The quality and independence of reference data from the map are more important than their quantity.
- **Description of accuracy assessment methodology should be detailed; reference data should be publicly available and include comprehensive metadata.** Key reporting elements are outlined in section 2.2, and Tables 3.1, 4.1 and 5.1.
- **The area of land cover classes should be estimated from the reference sample** and not derived via map pixel counting (section 5.2). The same reference sample can be used for both map accuracy assessment and land cover class area estimation.
- In addition to the opening of medium-resolution satellite data archives, **new sources of reference data have emerged** (Chapter 6) since the previous protocol by Strahler et al. (2006), e.g., lidar (section 6.2) and data from unoccupied aerial vehicles (UAVs) (section 6.3). While reference data access and sources will likely further improve in the future, the currently available data sources reviewed in Chapter 6 will remain relevant for validation of historic time-series maps.
- **Future directions and challenges** in land cover map accuracy assessment (Chapter 7) include the need for more funding and further methodology development to facilitate operational validation updates (section 7.1), near-real-time accuracy assessment (section 7.2), and the production of standardized reference datasets (section 7.3). Local map quality metrics (section 7.4) will likely begin to supplement overall and class-specific accuracy metrics estimated from the reference sample (section 2.7), as a response to the needs of the user community.
- The **Appendix** presents recent examples of national- to global-scale map accuracy assessment efforts.

1. Introduction

There have been significant changes in the amount of Earth observing data available and advances in technology and computing capabilities since the Land Product Validation (LPV) subgroup published its first protocol for the validation of global Land Cover products nearly 20 years ago (Strahler et al., 2006). The LPV subgroup was established in the year 2000, in the Earth Observation System (EOS) era (Morisette et al., 2002). The LPV subgroup falls within the Working Group on Calibration and Validation (WGCV), one of five Committee on Earth Observation Satellites (CEOS) working groups. This document addresses efforts within the Land Cover focus area of the LPV subgroup (https://lpvs.gsfc.nasa.gov/LandCover/LC_home.html) along with nine other thematic focus areas that correspond to essential climate and biodiversity variables. The LPV subgroup relies on voluntary contributions of the experts from each thematic focus area, with 2-3 co-leads at a time serving to coordinate the activities. Providing and updating community good practice guidelines on the validation of the satellite-derived map products is a primary objective of each focus area.

1.1 Scope of the guidelines

This document is aimed at updating the previous CEOS LPV global land cover validation guidelines, published in 2006 (Strahler et al., 2006). The main focus of the update is to provide an **overview of the land cover validation literature** published since, the novel sources of reference data (e.g., airborne lidar, high frequency very high-resolution optical data) in relation to the validation methodology, and the emerging issues in accuracy assessment, such as the need for operational validation updates and near real-time map validation. Like the original document, these guidelines are primarily focused on validation of **global- to continental-scale maps**, and as such will not address all the regional specifics of local-scale validation efforts. We are discussing general validation principles for **multi-class and single-class land cover maps**, without focusing on the validation specifics (e.g., definitions or response design) of each individual land cover theme. In addition, we cover the topic of **land cover class area estimation** based on the reference sample classification, which goes hand-in-hand with map accuracy assessment but was not covered by the original guidelines (Strahler et al., 2006).

As high-level community guidance, this document is meant **to be read prior to initiating a particular mapping project**, as the validation needs to be planned before map production is started, and not as an afterthought. We point map producers to the relevant publications and textbooks with detailed descriptions of the methods to be implemented. The current document is not meant to provide an end-to-end, step-by-step manual for every validation project. Instead, we provide **an overview of methods** of land cover map validation and area estimation **for a wide variety of use cases**, i.e. validating

pixel- and polygon-based maps using various sampling and response designs dictated by the nature of the maps and the type of reference data used, assuring the quality of the reference data, and selecting an appropriate estimator in each case. We also provide information to land cover map users that will enable them to evaluate the maps based on the reported accuracy results and to select the most appropriate map for their specific application.

The overarching goal of the guidelines is to **increase awareness** within the land cover mapping community of the complexity and constant development of the validation methods and issues and to **encourage correct implementation** of the validation methods. To do this, we have identified, within every thematic section, the potential limitations and caveats of each method, noting that there is no universal solution or a single 'best practice' approach to validation, but rather a variety of methods, each variably suited for a particular application. Section 2.2 lists the overarching principles of accuracy assessment and points to various places in the document providing detailed explanation of the presented concepts. We also **emphasize the importance of documenting** the validation methodology that has been implemented in each case **in a transparent and standardized way** (see Tables 3.1, 4.1 and 5.1).

1.2 The CEOS LPV validation stages

For a satellite-derived product to be considered a reliable source of information, it is required to have its accuracy assessed by comparison with an independent reference dataset. This process is also referred to as 'validation' (Justice et al., 2000), and both terms ('accuracy assessment' and 'validation') are used interchangeably throughout this document. At the CEOS validation stage 1 (Figure 1.2.1), the reference data sample size could be small and not necessarily a probability sample of the entire map. Further, the reference data might not cover the entire mapping time period. Publication of the validation results in the peer-reviewed literature is not required for this validation stage, and for land cover mapping specifically, such validation would typically not be sufficient to publish a map product in a peer-reviewed journal. Still, validation stage 1 provides some information regarding the possible sources and location of errors in the map being evaluated.

Validation Stages - Definition and Current State		Variable
0	No validation. Product accuracy has not been assessed. Product considered beta.	
1	Product accuracy is assessed from a small (typically < 30) set of locations and time periods by comparison with in situ or other suitable reference data.	**Snow** **Fire Radiative Power**
2	Product accuracy is estimated over a significant (typically > 30) set of locations and time periods by comparison with reference in situ or other suitable reference data. Spatial and temporal consistency of the product, and its consistency with similar products, has been evaluated over globally representative locations and time periods. Results are published in the peer-reviewed literature.	**fAPAR** **Phenology** **Biomass** **Evapotranspiration**
3	Uncertainties in the product and its associated structure are well quantified over a significant (typically > 30) set of locations and time periods representing global conditions by comparison with reference in situ or other suitable reference data. Validation procedures follow community-agreed-upon good practices. Spatial and temporal consistency of the product, and its consistency with similar products, has been evaluated over globally representative locations and time periods. Results are published in the peer-reviewed literature.	**LAI** **LST & Emissivity** **Active Fire** **Burned Area** **Vegetation Indices**
4	Validation results for stage 3 are systematically updated when new product versions are released or as the interannual time series expands. When appropriate for the product, uncertainties in the product are quantified using fiducial reference measurements over a global network of sites and time periods (if available).	**Land Cover** **Albedo** **Soil Moisture**

Figure 1.2.1 The CEOS LPV validation stages. 'Variable' corresponds to the CEOS LPV focus area and indicates the maximum validation stage currently reached by at least one of the products of that focus area. Source: https://lpvs.gsfc.nasa.gov/, accessed on September 4, 2025.

At the CEOS validation stages 2 and 3, the reference sample size is typically larger than that for stage 1 (> 30 sample units/locations) and the reference sample is a probability sample of the population being mapped, with reference data available for the entire mapping time period. Both stages require publication of the validation methodology and results in the peer-reviewed literature, with peer review serving as an independent verification of the accuracy assessment quality. Hence, most of the global and continental-scale land cover maps published in reputable peer-reviewed journals have been validated to stage 2 or 3. The difference between these stages is that stage 2 validation, while being sufficiently detailed, does not necessarily adhere to the community-agreed-upon good practices, i.e. some aspects of the validation might not have been performed properly or the description of methodology was not sufficiently detailed (Figure 1.2.2).

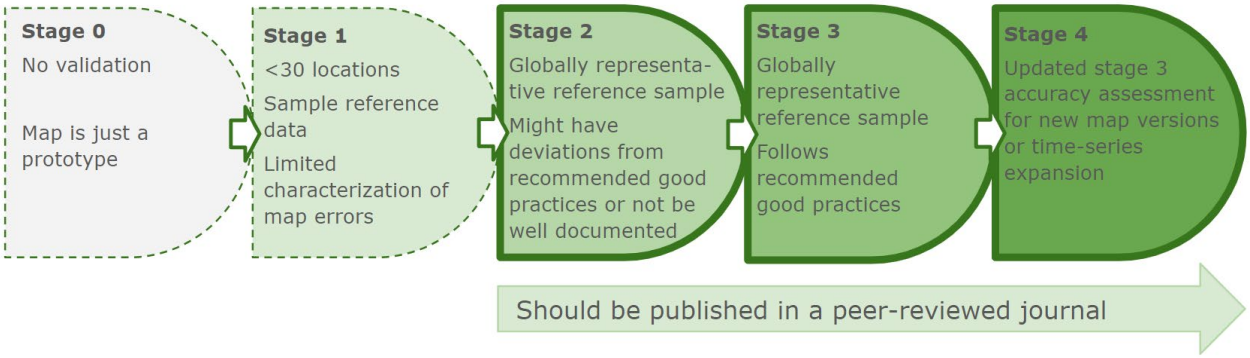

Figure 1.2.2 Key characteristics of each CEOS validation stage in the context of land cover mapping. At least stage 2 (and ideally stage 3) is required for publishing a land cover map in a peer-reviewed journal, and at the same time a peer-reviewed publication of a validation methodology is a requirement for reaching validation stage 2 and above. The recommended approach is to aim for validation stage 3 for newly created maps (and then stage 4 if the map is being updated). Please note that these validation stages present the variety of validation levels of existing and newly created maps, and not the sequential stages required for each map to go through. Also note that there is no standard definition of what 'globally representative' sample of locations means, although this term is used in the definition of the CEOS validation to distinguish Stage 2 and above from Stage 1. The rest of this document does not use the term 'globally representative sample' (see section 2.1 of Stehman and Foody (2019) for more discussion on the subject), instead the focus is on probability sampling.

Stage 3 validation, as defined by CEOS, uses the term 'uncertainties', but for land cover mapping 'errors' should be used instead, as stage 3 does not require per-pixel uncertainty modeling (propagation). Section 2.6 of the current protocol describes the difference between map accuracy and model/algorithm/classification uncertainty, and section 7.4 provides an overview of the methods of local map quality assessment (based on classification uncertainty metrics and spatial interpolation of accuracy metrics obtained from an independent reference dataset). While per-pixel uncertainty and accuracy metrics might become a standard request from the user community (especially the climate modelers) in the coming years, the current protocol's core recommendations and CEOS validation stages only require estimation of the overall (also referred to as 'global') and class-specific accuracy metrics (section 2.7). Both types of recommended metrics quantify the accuracy within the entire mapping region or sub-region, with overall metrics (e.g., overall accuracy) providing a single accuracy measure for all classes in the map, and class-specific metrics (e.g., producer's and user's accuracy) quantifying omission and commission errors of each mapped land cover class. "***Errors*** *in the product and their associated structure are* ***well quantified***" from the Stage 3 definition (from Figure 1.2.1 with '*uncertainties*' replaced with '*errors*') then means that the standard errors (or confidence intervals) of the estimated overall and class-specific accuracy metrics are provided along with the estimates (section 5.1, Table 5.1).

Validation stage 3 is a 'gold standard' validation (from a statistical perspective) for a given product version. When land cover maps are being updated, either due to the algorithm updates or expansion of the map time-series for the new year or back in time, the validation needs to be updated, thus moving the map product to validation stage 4. Land cover mapping is marked as having reached validation stage 4 in Figure 1.2.1, because there have been precedents of land cover maps reaching validation stage 4 (e.g., Tsendbazar et al., 2021, and section 7.1 of the current document). This does not mean that stage 4 validation is a common practice in the land cover community yet, although several global land cover monitoring programs are moving towards incorporating operational validation updates into their workflow. One of the major hurdles for shifting to validation stage 4 is lack of dedicated funding for operational validation updates of land cover maps. Collecting and using high quality and up-to-date reference data to assess the accuracy of land cover information produced from Earth Observation data is of fundamental importance to inform map users. This is particularly important now that land cover maps can be produced with increasing ease with high temporal frequencies for which the map validation is lagging behind. See section 7.1 for the methodological details on the implementation of validation stage 4.

1.3 Current state of global and continental-scale land cover and change mapping and validation

Land cover and change maps are often generated by classifying imagery into thematic classes. The increased access to large amounts of high-quality global remotely-sensed data in the last two decades has enabled individuals across the globe to both map and monitor global land cover. Since the previous protocol (Strahler et al., 2006), image providers of medium-resolution (10+ m) remote sensing data have largely addressed prior issues of accessibility and equitable access to Earth resource monitoring data (Jenice and Raimond, 2015). Image providers are moving towards processing satellite data to the analysis-ready level (Dwyer et al., 2018; Potapov et al., 2020) and harmonizing data streams from multiple satellite sensors (Claverie et al., 2018; Frantz, 2019). An increasing amount of satellite data are hosted by cloud-based image processing platforms (Ferreira et al., 2020; Gorelick et al., 2017), which simplifies utilizing data from multiple sensors for land cover and change mapping (Fortin et al., 2020) and makes access to computational resources less of a limiting factor when making global land cover and change maps (Wulder et al., 2018). Now, effective use of imagery to produce high quality maps is limited mainly by the map producers' (1) familiarity with principles of remote sensing, including the limitations and processing artifacts of the satellite data; (2) understanding of land cover class definitions and the feasibility of discerning a particular land cover class in the given data; (3) access to existing training data or the ability to create training data for machine learning algorithms; (4) familiarity with the algorithm parameters; (5) familiarity

with global land cover and ability to identify errors and to guide map iterations; and (6) understanding of validation principles, including the quality of reference data (Saah et al., 2020; Wulder et al., 2018).

Map validation and intercomparison activities have increased over the past two decades, including statistically rigorous accuracy assessment of global land cover (GLC) maps. 'Statistically rigorous' accuracy assessment (see section 2.1.1) is defined as relying on a probability sample, estimators (formulas) consistent with sampling design, and having standard errors of the estimates provided along with the accuracy estimates (Stehman, 2001; Stehman and Foody, 2019). Stehman and Foody (2019) report an exponential increase in the number of land cover-focused articles published in *Remote Sensing of Environment* (RSE) that include accuracy assessment in the last 50 years. This is likely related to an increasing number of articles focused on the methodology of map accuracy assessment being published in RSE and other remote sensing-focused journals (e.g., Congalton, 1991; Olofsson et al., 2014, 2013; Stehman, 2013, 2009; Stehman and Czaplewski, 1998; Stehman and Foody, 2019, etc.), which improved awareness of the importance of quantitative map accuracy assessment in the land cover community and enabled practical applications. The CEOS requirement that a land cover map be accompanied with a rigorous accuracy assessment has been implemented by several key global initiatives (Herold et al. (2016), Table 1.4). With increased availability of high (1-10m) and very high (<1m) resolution and timely remote sensing data, more sample sites have been used for assessing map accuracies (e.g., more than 30,000 sample sites in Chen et al. (2015) and Gong et al., (2013)). Crowdsourced reference data collection efforts (section 6.5) and improved reference data collection tools (e.g., Collect Earth Online, Geo-Wiki) have enabled an increase in the potential number of sample sites used for accuracy assessment, although the number of sample sites does not necessarily translate into the increased quality of reference data (see section 4.2). Sample size planning is recommended to calculate the minimum sample size necessary to achieve desired precision of the estimates (see section 3.5).

Several land cover maps have been validated to stage 3, which is a recommended validation stage for a given product version (section 1.2). Stage-3 validated maps (Table 1.4) include single-class land cover and change maps, such as forest cover, cropland or surface water dynamics (Hansen et al., 2013; Pickens et al., 2020; Potapov et al., 2022b), as well as multi-class maps (Defourny et al., 2017; Gong et al., 2013; Hansen et al., 2022; Potapov et al., 2022a; Zhang et al., 2021). The Copernicus global land cover monitoring efforts have implemented a stage 4 global land cover validation system as part of providing data as global service (Defourny et al., 2020; Tsendbazar et al., 2021). Global tree cover extent produced by the University of Maryland Global Land Analysis and Discovery (UMD GLAD) team from a time-series of tree height maps has been validated to stage 4, as the maps produced using the same model were validated first for 2019

(Potapov et al., 2021), and then for 2000 and 2020 (Potapov et al., 2022a). For a more complete list of global land cover mapping efforts and their validation stages, see section 1.4 below and Table 1.4.

Many global land cover maps, however, are still being produced and made available without an independent, rigorous land cover validation effort. This includes comparisons being done without a probability sample of quality reference data or with only a limited sample (e.g., Venter et al., 2022), accuracy estimates that lack transparency (no detailed method description or published reference data, see Table 1.4), and the use of methods that cannot be reproduced. The issue is amplified by increasing processing capacities and use of Artificial Intelligence (AI) that enable generating maps much quicker and easier and can lead to an inflation of land cover information in the absence of rigorous accuracy assessment. There is a risk that such products are being marketed to the users from a qualitative 'it looks good' perspective or with a promise of frequent updates, while critical issues with the map's quality are not explained or quantified. This results in potential misuse of the data and invalid conclusions being drawn from the maps, e.g., when classification errors in two maps are being interpreted as land cover change (Bontemps et al., 2012). In addition, there have been limited efforts to assess the quality of land cover change estimates globally using unified reference databases (Olofsson et al., 2012); potential benefits of such unified datasets and challenges of their creation are discussed in section 7.3.

Even when various independently produced land cover maps are validated following good practice guidance (Olofsson et al., 2014; Stehman and Foody, 2019) and the accuracies of individual maps are reasonable, there may still be significant differences in the extent of land cover classes derived from these individual maps. Comparative validation of these maps, based on a probability sample (also referred to as 'benchmarking', see section 2.8), using stratification to target the areas where these maps disagree (Lamarche et al., 2017), is then suggested to reconcile these differences and move towards consensus on land area change. The most recent state-of-the-art comparative validation of 10 m global land cover maps (Xu et al., 2024) revealed substantial differences between the maps' ability to represent the spatial detail of land cover. This kind of validation effort helps to demonstrate whether the promise of remotely sensed mapping to vastly improve the understanding of the Earth's dynamics (Townshend et al., 1991) has been successfully realized in the last decades.

1.4 Global land cover maps

Since the first iteration of this protocol (Strahler et al., 2006), the number of land cover map products has increased dramatically (Radeloff et al., 2024; Song, 2023; Woodcock and Ozdogan, 2012). Today, users have access to a variety of maps,

differentiated by their spatial, temporal, and thematic resolution. Coarser resolution maps (500 m - 0.05 degrees) are best suited for climate and dynamic vegetation modeling and other applications that might prioritize annual updates (e.g., MCD12Q1 in 2001-2021, Friedl and Sulla-Menashe (2022)) and long historic time-series (e.g., back to 1982 with AVHRR, Song et al. (2018)) over spatial resolution. Finer resolution maps (10m - <500m) are more suitable for various land management applications requiring higher spatial detail but are not necessarily updated annually or have historic coverage beyond the past two decades.

It is also important to distinguish multi-class land cover maps with generalized global legends (Brown et al., 2022; Buchhorn et al., 2020a, 2020b; Chen et al., 2015; Defourny et al., 2017; Friedl et al., 2022; Friedl and Sulla-Menashe, 2022; Gong et al., 2013, 2019; Hansen et al., 2022; Karra et al., 2021; L. Liu et al., 2020; Potapov et al., 2022a; Zanaga et al., 2021, 2022; Zhang et al., 2021) from single-class legend maps, focusing on a particular land cover theme, such as forest (Bourgoin et al., 2024; Feng et al., 2016b; Hansen et al., 2013), surface water (Feng et al., 2016a; Lamarche et al., 2017; Pekel et al., 2016; Pickens et al., 2020), cropland (Potapov et al., 2022b; Thenkabail et al., 2021), urban extent (X. Liu et al., 2020; Pesaresi and Politis, 2023; Schneider et al., 2009), bare ground (Ying et al., 2017), or mangroves (Giri et al., 2011). Some of these maps include land cover change mapped and validated directly, others provide annual land cover maps. In the latter case, land cover changes can be derived using post-classification comparison of land cover maps from two or more dates (see section 2.5), although it is not recommended, as overlapping errors in the annual maps may exceed real land cover change in magnitude. Regardless of whether land cover change is mapped directly or derived from map comparison, the land cover change class should be validated, and its area for reporting purposes should be estimated from the reference sample. Chapter 2 of the current document provides further discussion regarding land cover, land use and land cover change maps, as well as categorical maps vs. continuous fields of land cover classes (e.g., % vegetation cover or built-up surface).

When comparing or harmonizing land cover maps, both multi- and single-class, it is important to harmonize the legends of these maps (Vancutsem et al., 2012), as differences in employed land cover class definitions can lead to substantial differences between the maps. García-Álvarez et al. (2022) illustrated the diversity of legends currently employed for global multi-class land cover maps. While coming up with a single standardized land cover legend is often not feasible or practical, varying land cover class definitions can be used in different maps as long as these definitions are clear, comprehensive, and consistently employed across the globe within each map (Potapov et al., 2023).

The CEOS LPV Land Cover (LC) focus area maintains a periodically updated list of single- and multi-class continental- and global-scale land cover and change maps of

all spatial resolutions (https://lpvs.gsfc.nasa.gov/producers2.php?topic=LC) that have high-quality published validation information. The current CEOS LPV LC co-leads are tasked with evaluating newly published datasets to confirm that the datasets included in the list are validated to at least Stage 2 (preferably - Stage 3). Factors such as the lack of description of the validation methodology or sampling design flaws might lead to the dataset being excluded from the list. Please contact the current LC focus area leads if you notice any omissions in the dataset list.

Table 1.4 lists some of the global multi-class land cover maps with generalized global legends that have the resolution of 500m or finer and that have been updated in the past 10 years, along with metrics characterizing the quality of the map validation. These metrics include presence and completeness of validation information, availability of validation data and validation updates. Data users are encouraged to review the validation information of the newly published maps, including sampling design, response design and analysis components, to ensure that the map validation follows the good practice guidance from this document (Chapters 3, 4 and 5, respectively) and other publications (Olofsson et al., 2014; Stehman and Foody, 2019). If the sample unit-level validation data (reference data and reference labels for each sample unit) are provided along with the map, as recommended by these good practice guidance documents, data users are encouraged to review it. If sample labels do not seem to be reproducible or of high-quality judging from the provided reference data, data users should inform the map providers and/or the editors of the peer-reviewed journals that published the map. In case the detailed validation information is not publicly available, data users should be able to request it from the data providers. Increased transparency and level of detail of validation information should be a priority of the global land cover mapping community, as it will help build confidence in the land cover and change information provided by these global land cover maps.

1.5 Requirements for Land Cover Essential Climate Variable (ECV)

Ensuring high quality of data products for the purposes of climate monitoring (Bontemps et al., 2012) is one of the main purposes of the WGCV-LPV activities, including land cover and land use change products discussed in the current protocol. The Global Climate Observing System (GCOS) formulated requirements for the data products to support characterization of the Land Cover Essential Climate Variable (ECV) (WMO, 2022). Three distinct data products related to Land Cover ECV are identified in the 2022 GCOS Implementation Plan (Table 1.5):

- ***Land Cover***, defined as the observed (bio)-physical cover on the Earth's surface for regional and global climate applications;

- ***Maps of High-Resolution Land Cover***, defined as the observed (bio)-physical cover on the Earth's surface for monitoring changes at local scales (suitable for adaptation and mitigation);
- ***Maps of Key IPCC Land Classes, Related Changes and Land Management Types***, defined as land cover classes to be used for the estimation of GreenHouse Gas (GHG) emissions and removals following the Intergovernmental Panel on Climate Change (IPCC) guidelines.

For each of these data products, a set of requirements and their threshold values are identified for various maturity levels (Goal, Breakthrough and Threshold, Table 1.5). In terms of ***spatial resolution***, 2 out of 3 Land Cover ECV products are at the Goal level (Table 1.4). For the 'Maps of High-Resolution Land Cover' to reach the Goal spatial resolution finer than 10 m, global time-series of <10 m optical data need to be publicly available, which is not yet the case (see section 6.1).

In terms of ***temporal resolution***, most existing global land cover maps are at the Breakthrough (1 year) level, with the exception of the Dynamic World map (Brown et al., 2022), which is updated in near-real-time, satisfying the Goal requirement. Dynamic World, however, should be used with caution, because it has low accuracy in heterogeneous landscapes (Xu et al., 2024), lacks quality assessment of seasonal and annual changes, and classifies each satellite image separately, which introduces inconsistencies on the boundaries between images. Copernicus Land Monitoring system has planned the production of sub annual land cover and forest monitoring products, increasing the portfolio of products at sub annual temporal resolution (https://land.copernicus.eu/en/news/lcfm-a-new-chapter-in-global-land-cover-monitoring). The Goal temporal resolution of 1 month poses thematic challenges of defining land cover and change. For example, for mapping the extent of seasonally variable landscapes, such as seasonally inundated forests and grasslands, with a monthly temporal resolution would require comparing the current month with a baseline of the same month from several prior years to identify whether there is a change in land cover.

The Goal requirement of ***timeliness*** is satisfied by operational near-real-time forest disturbance maps (Hansen et al., 2016; Nagatani et al., 2018; Reiche et al., 2021; Shimabukuro et al., 2007), which are typically updated with a delay under 1 month. The global forest loss map (Hansen et al., 2013) is updated within 3-6 months from the end of the year, which is approaching the Goal requirement. For the multi-class land cover maps (except the Dynamic World discussed above), the timeliness is at the Breakthrough level (1 year) or below. Updating global maps that have annual temporal resolution requires processing satellite data archives for the entire year and extensive computational resources, which makes meeting the Goal timeliness requirements challenging. At a minimum, systematic visual map quality checks (see section 2.10) need to be performed

before releasing the map to the users, or, ideally, - operational validation updates (see section 7.1), which decreases the timeliness of the map products, but contributes to the increased user confidence in map quality.

Temporal extent of the existing land cover products (Table 1.4) is limited by the length of the satellite image archives. Medium resolution (10 - 500 m) optical satellite data have consistent global coverage only from 2000 forward, starting with the global acquisition strategy of Landsat 7 and the launch of MODIS instrument onboard Terra satellite. To satisfy the Goal requirement of 30+ year temporal extent, public medium-resolution Earth observation missions such as the Sentinel and Landsat programs, need to have continued funding in the coming decades, to ensure continuity of data archives.

Required measurement uncertainty in the context of land cover mapping corresponds to required map accuracy or maximum tolerated map errors. In the requirements document it is expressed as a 95% confidence interval (CI) of the reported overall accuracy and commission and omission errors (user's and producer's accuracy) for each land cover class, i.e., the extent of those errors. For example, user's accuracy of 85% ± 5% (95% CI) would correspond to the 15% ± 5% (95% CI) commission error or 25% when expressed as a width of 95% CI of commission error (units of Table 1.5). The Threshold requirement is the maximum acceptable commission/omission error for individual land cover classes (WMO, 2022). In the existing high-resolution global land cover maps (Xu et al., 2024), most classes are at least at the Threshold accuracy level, with some classes, such as the extent of surface water, trees and bare ground reaching Breakthrough accuracies, and only perennial snow and ice class reaching Goal accuracy. "*An independent accuracy assessment using statistically robust, global or regional reference data of higher quality is required for any ECV land cover product*" (WMO, 2022) regardless of the achieved accuracy level.

Stability (change per decade in a 95% confidence interval of reported accuracy metrics, i.e., decadal change of map accuracy) will need to be assessed once multi-decadal land cover map accuracy assessment efforts are available. Tsendbazar et al. (2021) proposes metrics to assess the stability of the accuracy of annual global land cover maps.

The World Meteorological Organization (WMO) is currently reviewing ECVs and requirements for data products for their measurement and monitoring (https://wmo.int/media/magazine-article/call-review-of-essential-climate-variable). An updated set of requirements is planned to be published with a new GCOS Implementation Plan in 2028.

Table 1.4 Global categorical land cover maps with generalized multi-class legends with spatial resolution of 500m and finer, which map land cover for at least one year after 2013, along with the indicators of the quality of the validation information, published along with the map. MODIS stands for Moderate Resolution Imaging Spectroradiometer; MERIS stands for Medium Resolution Imaging Spectrometer, PROBA-V stands for PROBA-Vegetation, HJ-1 stands for Huan Jin-1. USGS stands for United States Geological Survey, ESA - European Space Agency, UMD GLAD - University of Maryland Global Land Analysis and Discovery laboratory. Links accessed on September 4, 2025.

Sensor/ satellite	Land cover map	Resolution (m)	Temporal Frequency and Range	Link to data	Link to documentation/ publication	Validated (Yes/No)	Validation methodology well described (Yes/No)	Validation data available (Yes/No)	Validation updated (Yes/No)	Validation stage
MODIS	MCD12Q1 Land Cover Type	500	Yearly, 2001-2021	USGS	User guide Validation status	Partially	Yes Olofsson et al., 2012; Stehman et al., 2012	No	No	Stage 2
MERIS	ESA CCI Land Cover	300	Yearly, 1992-2015	ESA	User guide	Yes	Yes Link to description GlobCover 2009 validation dataset	No	No	Stage 3
PROBA-V	Copernicus GLS-LC100	100	Yearly, 2015-2019	Copernicus	Buchhorn et al., 2020	Yes	Yes Validation report	No	Yes Tsendbazar et al., 2021	Stage 4
Landsat	UMD GLAD Global Land Cover and Land Use Change	30	Bi-decadal change, 2000-2020	UMD GLAD	Potapov et al., 2022	Yes	Yes	Partially: Cropland, Surface water	Partially, only forest extent (initial validation Potapov et al., 2021)	Stage 3, Forest extent Stage 4
Landsat	UMD GLAD Global Land Cover and Land Use	30	One year, 2019	UMD GLAD	Hansen et al., 2022	Yes	Yes	Yes Link	No	Stage 3
Landsat	GLC-FCS30	30	One year, 2015	Liu et al., 2020	Zhang et al., 2021	Yes	Yes, but not a probability sample, compilation of multiple existing datasets	Yes Link	No	Stage 3

Sensor/ satellite	Land cover map	Resolution (m)	Temporal Frequency and Range	Link to data	Link to documentation/ publication	Validated (Yes/No)	Validation methodology well described (Yes/No)	Validation data available (Yes/No)	Validation updated (Yes/No)	Validation stage
Landsat	FROM-GLC30	30	Intermittent, 2010, 2015, 2017	Pengcheng Laboratory, 2015 and 2017	Gong et al., 2013	Yes	Yes	No	No	Stage 3
Landsat, HJ-1	Globeland30	30	Intermittent, 2000, 2010 2020	Globeland30	Chen et al., 2015	Yes	No, Missing sample interpretation protocol description	No	No	Stage2/3
Sentinel-2	FROM-GLC10	10	One year, 2017	Pengcheng Laboratory	Gong et al., 2019	Yes	Not clear which validation dataset used, no details except overall accuracy number	Perhaps this dataset	No	Stage 2
Sentinel-2	ESRI Land Cover	10	Yearly, 2017-2022	ESRI	Data description Karra et al., 2021	Yes	No, lacking sampling design and sample interpretation protocol details	No	No	Stage 2
Sentinel-1, Sentinel-2	WorldCover (ESA)	10	Yearly, 2020, 2021	ESA Zanaga et al., 2022 (v200) Zanaga et al., 2021 (v100)	User manual v100 (2020) and v200 (2021)	Yes	Yes Validation report v100 (2020) Validation report v200 (2021)	No	Yes Tsendbazar et al., 2021	Stage 4
Sentinel-2	Dynamic World	10	Near Real Time, 2017-Present	Dynamic World	Brown et al., 2022	Yes	Response design described in detail, sampling design and analysis not clear	Yes, Brown et al., 2021	No	Stage 2/3

Table 1.5 Requirements for the Land Cover Essential Climate Variable (ECV) data products from the Global Climate Observing System (GCOS) 2022 Implementation plan (WMO, 2022). Vertical resolution requirement is omitted from the table, because it does not apply to the Land Cover ECV products. *Goal* is an ideal requirement above which further improvements are not necessary; *Breakthrough* is an intermediate level between threshold and goal which, if achieved, would result in a significant improvement for the targeted application; *Threshold* is the minimum requirement to be met to ensure that data are useful. *In the context of land cover mapping, 'required measurement uncertainty' should be interpreted as 'required map accuracy' or 'maximum tolerated map errors'. Confusion between the terms 'uncertainty' and 'errors' is also discussed in section 1.2.

Requirements	Requirement levels	ECV Product		
		Land Cover	**Maps of High-Resolution Land Cover**	**Maps of Key IPCC Land Classes, Related Changes and Land Management Types**
Horizontal (spatial) resolution	Goal	100 - 300 m	< 10 m	10 - 300 m
	Breakthrough	300 m - 1 km	10 - 30 m	300 m - 1 km
	Threshold	> 1 km	300 - 100 m	1 km - 1 degree
Temporal resolution	Goal	1 month		
	Breakthrough	1 year		
	Threshold	5 years		
Timeliness (reporting/processing delay)	Goal	3 months		1 month
	Breakthrough	1 year		
	Threshold	5 years		
Temporal extent (time span)	Goal	> 50 years	30 - 50 years	> 100 years
	Breakthrough	10 - 50 years	10 - 30 years	50 years
	Threshold	One time		30 years
Required measurement uncertainty* (95% confidence interval of overall map accuracy, omission and commission errors of individual land cover and change classes, and of area estimates)	Goal	5%		5%
	Breakthrough	20%		15%
	Threshold	35%		25%
Stability (change per decade in 95% confidence interval of omission and commission errors of individual land cover and change classes, i.e., decadal change of accuracy of individual map classes)	Goal	5%		
	Breakthrough	15%		
	Threshold	25%		

2. Definitions and general principles

2.1 Definitions

2.1.1 General terms

Design-based inference - protocol for generalizing from a sample to a population in which properties of estimators such as bias and variance are determined based on the randomization distribution (e.g., frequency distribution, histogram) of the estimator over the set of possible samples for the sampling design implemented. A probability sample is required to implement design-based inference. A statistically rigorous map accuracy assessment (see definitions below) in a design-based inference framework is a core recommendation of the current document.

Model-based inference - protocol for generalizing from a sample to a population in which the observations on each element of the sample are assumed to have been generated from a model; properties such as bias and variance are dependent on the specification of the model, and it is critical to assess whether the assumed model is tenable given the data. Model-based inference is conditional on the sample selected (i.e., it does not consider variation over the possible samples that might have occurred) and does not require a probability sample.

Accuracy assessment - process of quantitatively assessing the quality of a map by comparing the map labels to independently derived reference labels that are the best practical determination of the ground condition. Map accuracy is assessed via estimating accuracy metrics (e.g., overall, user's and producer's accuracy, see definitions below) from reference labels over a sample of locations. This differs from the current metrological definition of accuracy (JCGM, 2012) and might differ from other definitions of accuracy used in other CEOS working groups.

Validation - the process of assessing, by independent means, the quality of data products (Justice et al., 2000). Although 'validation' might be used in a broader sense than 'accuracy assessment' in the literature (e.g., for assessing the quality of map products using opportunistically collected reference data or other maps), in this document ***'accuracy assessment' and 'validation' are used interchangeably***, referring to the statistically rigorous map accuracy assessment.

Statistically rigorous - for design-based inference, a statistically rigorous accuracy assessment is one in which a probability sampling design is implemented and the

estimators (formulas) are consistent with the sampling design; i.e., properly weighted to account for inclusion probabilities (Stehman, 2001). Stehman and Foody (2019) added a requirement of quantifying the variability of the accuracy and area estimates by reporting standard errors or confidence intervals to the criteria of statistically rigorous assessment.

Reference label or ***reference classification*** - best practical determination of ground condition (i.e., closest to the reference land cover class as determined via a ground survey that is feasible and affordable) for an assessment unit (see definition below). If a field survey is not feasible, the ground condition (land cover class) is often determined via visual interpretation of high resolution imagery.

Reference data or ***validation data*** - collection of reference labels for a sample of locations. For statistically rigorous accuracy assessment of land cover maps reference data is a collection of reference land cover labels for a probability sample.

Reference data source - remotely sensed or ground survey data used to determine reference labels (see Chapter 6).

Map label or ***map classification*** - land cover class label assigned to a specific map unit (see definition of a map unit below) in a process of creating a land cover map.

Training data - a set of labeled locations used to train a land cover classification model. Training data does not have to originate from a probability sample. In fact, active learning mode, when training data are added interactively between classification iterations in the areas where the model does not perform well, implies targeted collection of training data in those areas of poor performance, often resulting in a non-probability sample. Training data used for building the model should not be used for accuracy assessment of the resulting land cover map to avoid optimistic bias. See more discussion about the independence between the training and the reference data in sections 2.2.2 and 2.2.3.

Testing or ***cross-validation*** - evaluation of the model performance based on a set-aside of a training dataset that informs the choice between model types, tuning of the model parameters and features, and helps assess the classification uncertainty for different land cover classes (e.g., to identify whether additional training data needs to be collected). Model/classification uncertainty refers to the repeatability of classification results between model runs. Such evaluation is not considered a part of the map accuracy assessment as defined by the current guidelines (see section 2.6 for more details).

Pixel counting or ***map-based area estimate*** - an approach used to estimate the area of a class in a thematic map produced from remotely sensed data. In this approach, the area of a class is the number of pixels allocated to the class in the map multiplied by the areal extent of a pixel. Pixel counting is not a recommended approach of estimating land cover

class area for the purposes of reporting, as area estimates derived using this approach are affected by map errors (see 'Omission error' and 'Commission error' below).

Sample-based area estimate - estimate of area using the reference class labels obtained for a sample from the population of interest. Statistically rigorous (see definition above) sample-based estimates of land cover and change class areas along with their uncertainties are recommended to be reported instead of map pixel counts in scientific and working papers, governmental reports, and monitoring and accounting frameworks, such as forest inventories and climate change mitigation initiatives. See also 'Estimate' below.

Categorical land cover map – a geographical representation of land cover as a set of discrete categories. Each location in the map has a label that indicates its thematic class. For example, a conventional land cover map displays a nominal level class label (e.g., forest, grassland, water, etc.) for each location and, as a whole, represents the spatial distribution of the land cover mosaic in the mapped region.

Continuous land cover map or ***continuous field*** - a type of land cover map where each location is categorized as a range of values (e.g., % forest cover or % cropland within a pixel) rather than a discrete class. Continuous fields can be used to reflect heterogeneity of traits within the same general land cover (e.g., structure of forests characterized by % canopy cover and height maps) and to characterize finer-scale land cover mosaic (as opposed to categorical land cover maps of the same resolution) when the same location has proportions of various land cover classes expressed as continuous fields (e.g., a mixed-class 30 m pixel can have 50% forest, 40% cropland and 10% water).

Operational (monitoring/mapping) - regularly updated land cover and change maps suitable for continuous monitoring, as opposed to one-off map products covering limited points in time.

Operational validation updates (section 7.1) - regularly updating reference data and accuracy assessment results for each new map release or time series expansion of land monitoring products.

Near real-time (monitoring/mapping) - identifying particular land changes continuously and as quickly as possible. The specific definition of 'near real-time' can range from seconds to months depending on the application, as opposed to traditional land cover monitoring, where maps are typically updated annually. Challenges of validating near real-time maps are addressed in section 7.2.

2.1.2 Sampling design terms

Sampling design **-** the protocol by which the reference sample units are selected (Stehman and Czaplewski, 1998). In a broad sense, sampling design includes specification of the following elements (sections 3.1 - 3.5): sampling unit, sampling frame, probability sampling design, strata and clusters (if used), sample size and allocation to strata. In a strict sense, the term 'sampling design' is used to refer to the specific protocol by which sampling units are selected (see 'Probability sample' below and section 3.3).

Population - the entire spatial region (e.g., defined by the geographic boundaries of a country or a region) or collection of elements (units) of interest (e.g., all pixels in a map) for the specific study.

Inclusion probability - the probability that an element (unit) of the population is included in the sample.

Sampling frame - a list, map, or other specification of sampling units in the population from which a sample may be selected (Lohr, 2010).

Sample - a subset of a population, for example, a subset of spatial units or area. A sample is a collection of sampling units or sampling points.

Probability sample - a method for selecting a sample implemented using a random or chance mechanism such that the inclusion probabilities of the elements of the sample are known and the inclusion probabilities for all elements of the population are greater than zero.

Sampling unit - a unit that can be selected for a sample (Lohr, 2010).

Sample unit or sample pixel/point/polygon - a unit that has been actually selected for a sample. Please, note that it is incorrect to refer to the units in a sample as 'samples': this creates confusion between individual sample realizations (see 'Sample' above), which can be referred to as 'samples', and the units within each sample, which should be referred to as 'sample units' (or 'sample pixels/points/polygons'). Example of correct usage: "A simple random sample of 600 pixels was selected… Each sample unit (pixel) was interpreted…". Example of incorrect usage: "We selected 600 pixels via simple random sampling… Each sample was interpreted…"

Cluster sampling *(one stage and two stage)* - secondary sampling units (SSUs) are combined into groups and these groups (clusters) are the initial units (primary sampling units, PSUs) selected by the sampling design. For example, 30 m pixel SSUs can be grouped into 10x10 pixel clusters (PSUs). In one-stage cluster sampling, all SSUs within

each sampled PSU are observed (e.g., each 30 m pixel within a 10x10 pixel cluster), whereas in two-stage cluster sampling a sample of SSUs is selected within each sampled PSU (e.g., a simple random or systematic sample of 30 m pixels within a 10x10 cluster).

Stratification - a partitioning of the population elements into groups (strata) such that each element of the population belongs to one and only one stratum. Please note that 'stratum' is a singular form (one stratum), and 'strata' - plural (many strata).

Sample allocation - the distribution of the total sample size to strata (i.e., sample size per stratum).

2.1.3 Response design terms

Response design - the protocol defining how a reference class label is assigned, and how a decision on the agreement between the map and the reference class label is made (Stehman and Foody, 2019). Response design includes defining the spatial unit of the assessment (see 'Assessment unit' below), the sources of information used to determine the reference label, the labeling protocol, and the definition of agreement between the map and the reference classification (Olofsson et al., 2014).

Assessment unit - the spatial unit (e.g., point, pixel or polygon) that serves as the basis of the comparison between the map label and the reference label to determine agreement. If a sampling unit (see definition above) is a cluster of map units (e.g., block of map pixels), an assessment unit for map accuracy assessment should be equivalent to a map unit (i.e., assessment should be performed at the level of secondary sampling units, see definition of cluster sampling above). If a sampling and assessment unit is a point, a spatial support unit equivalent to a map unit should be employed to define agreement between the map and the reference classification. In the context of land cover class area estimation, the assessment unit is the basis for determining the reference land cover class.

Spatial support unit - the area that is taken into account when assigning reference class labels to the assessment unit. See introduction of Chapter 4 for more discussion about the relationship between spatial support unit, assessment unit and sampling unit.

Map unit - the smallest spatial unit in which map data is stored, e.g. a pixel or a polygon. Map units or their clusters often serve as sampling units (see definition above).

Map spatial support unit or ***Minimum mapping unit (MMU)*** - the area that is taken into account when assigning reference class labels to the map units. For example, a forest extent map can have a 30x30m map unit (pixel size), but employ the FAO definition of forest, which has an MMU of 0.5 ha, equivalent to 5.55 map units. That is, for a map unit

containing trees to be labeled as a 'forest' land cover class, it has to belong to a larger (at least 6 pixels) tree cover patch.

Majority interpretation approach - a sample labeling protocol in which a majority class label from multiple interpreters is selected as a final reference label for each assessment unit. This approach aims to reduce biases from the individual interpreters, and is often employed both in expert-based and in crowdsourced applications (see section 6.5).

Consensus interpretation approach - a sample labeling protocol in which interpreters work collaboratively to reach consensus regarding the final reference label for each assessment unit if the initial labels from different interpreters working independently differ. This approach allows identifying lower confidence cases, and focusing additional interpretation effort on those cases, thus attempting to reduce the interpreter variability instead of simply incorporating it into the estimated variance.

Field survey or ***ground survey*** - *in situ* data obtained via field visits that can be used to assess accuracy of land cover maps. Please note that the term 'ground truth' is sometimes used in the literature to refer to reference labels in a broad sense, whether derived on the ground or, for example, from high-resolution satellite imagery. In the current document we avoid using the term 'ground truth' because of this ambiguity, and because of the errors in reference data (see section 4.2) that are always present in the reference data (including field surveys) and contradict the 'truth' in the 'ground truth'.

2.1.4 Analysis terms

Analysis protocol – specification of the measures to be used to express accuracy and land cover class area and the procedures to estimate the selected measures from the sample (Olofsson et al., 2014).

Parameter - a number that characterizes a population (i.e., a census value). Examples of population parameters are the total area of a land cover class, or the area of the map that is correctly classified (see 'Overall accuracy' below) derived from a census of all population elements. True values of population parameters are usually impractical to measure and therefore unknown. In practice, the population parameters are estimated from the sample (see 'Estimate' below).

Estimator - a statistic that uses information in the sample to produce a value (number) that should be close to the population parameter. An estimator is the formula for calculating this statistic. An estimator is a random variable dependent on which sample units are selected.

Estimate or ***sample-based estimate*** - the actual value of an estimator obtained from the data for a particular sample selected.

Unbiased estimator - an estimator whose average, over all possible samples that could be selected by the sampling design implemented, equals the population parameter. An unbiased estimator could have estimates from some samples that are very far from the parameter but samples with overestimates are balanced out by samples with underestimates so that on average, the estimator equals the parameter.

Uncertainty/precision - repeatability of an outcome (e.g., values of an estimator over different samples). This agrees with a metrological definition of precision as "closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions" (JCGM, 2012). Uncertainty in this context is defined as the opposite of precision; i.e., low precision (low repeatability) means high uncertainty of the estimates. Uncertainty/precision of the sample-based estimates of the population parameters is characterized by their variance (see below, with high variance corresponding to low precision and high uncertainty). Model uncertainty refers to the repeatability of classification results between model runs (e.g. using different subsets of a training dataset).

Variance - the mean (over all possible samples) of the squared differences between the estimate for a sample and the population parameter; variance characterizes the repeatability of the estimator over the possible samples that could be selected by the sampling design implemented (e.g., repeatability of the estimated area of deforestation over all possible samples). Variance can be estimated from the one sample obtained. Because variance assesses repeatability of an estimator it is often characterized as a measure of uncertainty/precision.

Standard error - the square root of the variance of an estimator; the standard error characterizes variability on the same scale as the estimate (e.g., if estimating area in units of hectares, the standard error has units of hectares whereas the variance has units of hectares^2).

Confidence interval - estimates of lower and upper bounds (also referred to as 'confidence limits') for a parameter forming a range in which plausible values of that parameter lie; this interval takes into account variability due to sampling (i.e., an estimate produced from a sample is unlikely to match the population parameter). Assuming normal distribution of parameter estimates from different sample realizations, the 95% confidence interval is usually computed as ±1.96 x standard error.

Confusion matrix **or** ***error matrix*** – the cross-tabulation formed by the comparison of the predicted and actual class membership labels for a sample of cases (e.g., map pixels).

For map accuracy assessment, the predicted label is a map label (see definition above), and the actual class membership is a reference label (see definition above). The correctly labelled cases lie on the main diagonal of the matrix. Off-diagonal elements of the matrix highlight errors or misclassifications. See also Figure 2.7 and Table 5.1.1.

Overall accuracy - a proportion (or percentage) of the area that is correctly classified in a map. Overall accuracy can be derived as a sum of diagonal elements of a confusion matrix expressed in area proportions (see Table 5.1.1). Overall accuracy is an *overall accuracy metric*, because it is a single accuracy measure applied to the entire set of land cover classes in a map. Sometimes overall accuracy is also referred to as a *global accuracy metric*, even though it can be computed for the entire mapping region and its subregions or subdomains (e.g., overall accuracy of a global map can be computed for individual countries and biomes). To avoid confusion with the geographic scale at which these metrics are computed, in the current document we refer to the accuracy metrics characterizing the entire set of classes as 'overall accuracy metrics'.

Class-specific* or *per-class accuracy metrics - accuracy metrics characterizing the accuracy of individual land cover classes in a map, e.g., user's and producer's accuracy (see definitions below). Like overall accuracy, class-specific metrics can be computed for the entire map, or for its subregions.

Omission error – a misclassification viewed from the perspective of the actual class of membership. For example, a case (map unit) of class X may be incorrectly allocated to class Y and hence has been wrongly excluded (omitted) from class X. Map omission error for a single land cover class can be estimated from the confusion matrix and is often expressed as a proportion (or as percentage) of area incorrectly omitted from that land cover class in the map. See also 'Commission error' and Figure 2.7.

Producer's accuracy - is the accuracy of a classification for a single class from the perspective of a basic map maker (see Figure 2.7). It indicates how well the classification represents the class and is the complement of the omission error: producer's accuracy = 100% - omission error (%). In a binary classification, producer's accuracy of a target class is also referred to as 'sensitivity' or 'recall', and producer's accuracy of background class - as 'specificity' (see Table 5.1.2).

Commission error - a misclassification viewed from the perspective of the class of allocation. For example, a case (map unit) of class X may be incorrectly allocated to class Y and hence has been wrongly included (commissioned into) class Y. Map commission error for a single land cover class can be estimated from the confusion matrix and is often expressed as a proportion (or as percentage) of area incorrectly mapped as that land cover class. See also 'Omission error' and Figure 2.7.

User's accuracy - the accuracy of a classification of a single class from the perspective of a basic map user (see Figure 2.7). It indicates how often the class labelling will match what is actually observed on the ground and is the complement of the commission error: user's accuracy = 100% - commission error (%). In a binary classification, user's accuracy of a target class is also referred to as 'precision' (see Table 5.1.2).

F1 score - a metric often used to express the accuracy of a classification. It is calculated as the harmonic mean of the user's and producer's accuracies (see section 5.1).

2.2 Key principles of accuracy assessment

A perfect map would contain no error and thus be completely accurate. In reality, all maps contain errors, and it is important to know the nature and extent of these errors to ensure appropriate interpretation and use of a map. A map without an accuracy statement may be regarded as merely an untested hypothesis. Stehman and Foody (2019) proposed fundamental principles or good practices of map accuracy assessment as follows: the accuracy assessment should be (1) map relevant; (2) statistically rigorous; (3) quality assured; (4) reliable; (5) transparent, and (6) reproducible. In this section, we list specific aspects of accuracy assessment that fall within these six general principles and that are further elaborated on in various chapters of the current protocol. The protocol as a whole and the summary below are organized around the three major components of the accuracy assessment (Stehman and Czaplewski (1998), Figure 2.2): sampling design (Chapter 3), response design (Chapters 4 and 6), and analysis (Chapter 5). The list provided here is meant to serve as a broad guide to the key considerations when planning the map validation work. The specifics of sampling design, response design and analysis will be different depending on the objectives of the particular project and the map being validated, but the principles listed below should be followed for the accuracy assessment to be considered credible.

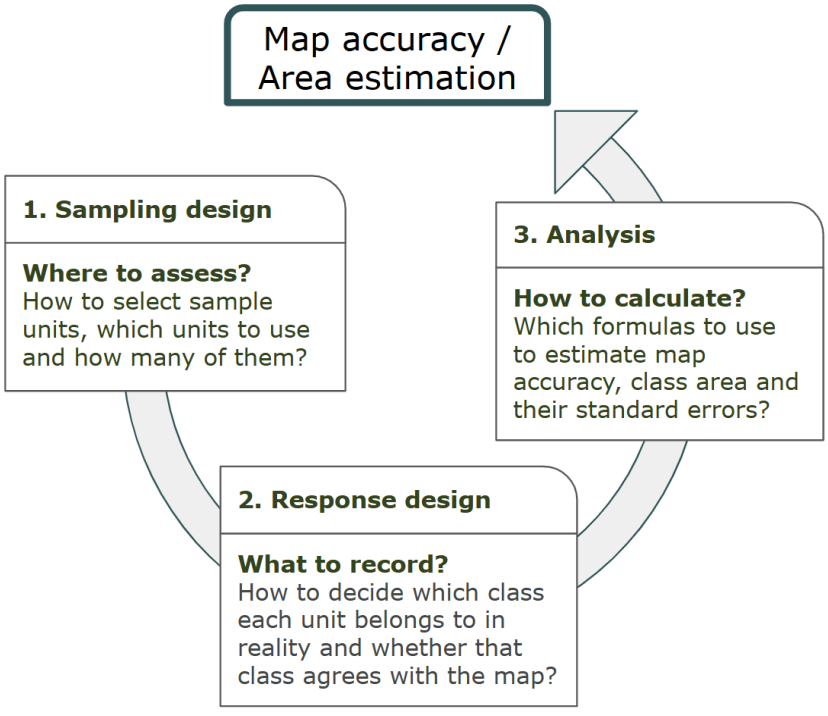

Figure 2.2. Main components of map accuracy assessment: sampling design (Chapter 3), response design (Chapter 4) and analysis (Chapter 5).

2.2.1 General principles

- *Accuracy assessment should be taken seriously.* Map accuracy assessment (as defined in sections 2.6 and 2.7) should be planned before map production is started, and a substantial budget separate from training data collection and other tasks should be allocated for reference data collection. Strahler et al. (2006) suggest that at least a third of the total mapping project cost and effort should be allocated towards validation.

- *Map accuracy assessment should be designed to fit the purpose.* See sections 2.3 - 2.8 of the current chapter for the overview of the diversity of land cover map types, accuracy assessment and area estimation objectives.

- *Design-based inference is recommended for map accuracy assessment.* This requires that a probability sampling design is implemented (Chapter 3), and the estimators for map accuracy and area depend on the selected sampling design (Chapter 5).

- *Detailed metadata, including sampling design information (Table 3.1), response design specifics (Table 4.1) and analysis approach (Table 5.1) should be reported to enable transparency and reproducibility of the accuracy assessment.* Documenting the validation protocol in a transparent and standardized way helps correctly interpret the reported map accuracy metrics and increases user confidence in the quality of the map.

2.2.2 Sampling design principles

- *Map accuracy assessment should be based on a probability sample of reference data to be statistically valid when using design-based inference.* Common probability sampling designs are discussed in section 3.3. Stratified random sampling is a recommended 'universally adequate' general purpose sampling design (Olofsson et al., 2014; Stehman, 2009a).

- *Information from a map, often the map that is being validated, can be used to enhance precision of the estimates via stratification.* Constructing the strata from the map that is being validated does not introduce bias in the accuracy estimators and does not violate the requirement of independence of the reference sample from the map, as the strata affect only precision of the estimates (Stehman and Foody, 2019). Stratification and its benefits are discussed in section 3.4.

- *Required sample size should be calculated depending on the objectives.* See section 3.5 for the overview of sample size planning and allocation among strata for different estimation objectives.

- *There is no need to geographically separate training and reference data if they are derived independently*, as there is no requirement that the same spatial unit cannot be included both in the training set and validation sample. Excluding geographic areas (polygons) that were used for model training and areas of their proximity (e.g., buffers around training polygons) from sampling will make the resulting estimates valid only for the areas outside of these excluded areas, and not for the entire mapped region. That is, the sample locations of training data are still part of the population that is being validated. An exception to this is when training and reference data are both derived from the same cluster sample (e.g., both come from a sample of high-resolution imagery or from field surveys); in this case it is advisable to use each cluster for training or validation (but not for both), to avoid potential bias of validation data being spatially closer to training data compared with the case when the training and validation data are sampled from the entire region of interest.

2.2.3 Response design principles

- *Reference data for validation should be separate from map training data.* Map training data do not have to originate from a probability sample (i.e. can be opportunistically collected), while reference data for validation has to be from a probability sample to allow use of design-based inference as described in Chapters 3 and 5. Reference sample data for accuracy assessment can be a set-aside of the training dataset, if the training sample is a probability sample, although this is not recommended, as biases/errors present in the training data may be transferred to the reference data resulting in potential overestimation of map accuracy (see section 4.2). If possible, reference data should be produced by interpreters different from map makers, and the process of reference data collection should be independent from training data collection (see Table 4.1). Experts producing reference data should not be aware of map labels for the sample locations.

- *Reference sample labels should be of higher quality than the map classification.* This is achieved either through obtaining reference data of higher quality (e.g., high-resolution imagery to validate medium-resolution maps) or by employing a more accurate process of deriving reference labels (e.g., manual determination of reference labels when the map is produced via automated classification) if using the same data for mapping and validation (Olofsson et al., 2014). See Chapter 4 for planning the protocol of reference data collection, and Chapter 6 for the sources of reference data.

- *Quality of reference data should be assessed.* See sections 4.2 and 4.3 for the factors affecting the quality of reference data and ways to account for these factors and incorporate the uncertainty of the reference data into the estimates.

- *Sample-unit-level reference data and reference labels (classification) should be provided for review to ensure transparency and enable independent verification of reference data quality.* If data privacy, ownership of intellectual property concerns preclude sharing the exact locations of sample units (e.g., field plots) or non-publicly-available reference imagery, the data should still be shared without geolocation information or in a form of non-georeferenced image previews with proper attribution/licensing information.

- *All deviations from the probability sampling design or reference data collection protocol should be carefully documented and reported.* For example, sample locations not visited in the field or assessed using lower quality reference data should be identified.

2.2.4 Analysis principles

- *Estimation formulas for accuracy and area should correspond to (be valid for) the selected sampling design*. Recommended estimators are unbiased, which means that these formulas account for inclusion probabilities associated with a particular sampling design. Chapter 5 refers to publications containing special case equations for various sampling designs.

- *Confusion matrix should be expressed in terms of area proportions (Table 5.1.1).* The sample-based estimates of the area proportions of the confusion matrix must take into account the sampling design (e.g., unequal inclusion probabilities in different strata when implementing a stratified sampling design).

- *User's and producer's accuracy or similar accuracy metrics quantifying omission and commission errors of individual land cover classes should be used in addition to overall accuracy.* This is particularly important when a target land cover class is small, and high overall accuracy might be misinterpreted. The use of the kappa coefficient of agreement is strongly discouraged by the existing good practice guidelines (Olofsson et al., 2014; Strahler et al., 2006). More information about map accuracy metrics can be found in sections 2.7 and 5.1.

- *If estimates of both map accuracy and area of target land cover class(es) are required, they can be derived from the same reference sample.* The sampling design could be chosen to increase precision of accuracy or area estimates based on the priorities of the specific project, but regardless of the selected design, the probability reference sample is valid to estimate both map accuracy and area of land cover class(es). Reporting area estimates in addition to map accuracy metrics usually requires very little additional work (applying area estimators to already produced sample reference data, see section 5.2), but is often not done by map

producers. Chapter 5 and section 2.7 provide further details on the relationship between area and accuracy estimation.

- *Variability of the accuracy and area estimates should be quantified by reporting standard errors or confidence intervals* (Stehman and Foody, 2019). In practice, standard errors of accuracy metrics (e.g., overall, user's and producer's accuracy) are often not reported, making it difficult for the users to assess whether the provided map accuracy estimates are reliable. Sections 5.1 and 5.2 provide references containing unbiased variance estimators for various sampling designs.

2.2.5 Additional points

- *Algorithm uncertainty should not be confused with map accuracy assessment.* Both provide useful information, but only the latter establishes correspondence between the map and an independently derived reference sample (see section 2.6 for more details).

- *Although the current recommendations for accuracy assessment are primarily based on design-based inference, model-based inference offers another statistically valid approach.* Model-based inference does not require a probability sampling design, and thus can be used with non-probability reference data (McRoberts et al., 2022; Stehman and Foody, 2019). However, the use of model-based inference for map accuracy assessment has been limited. Model-based methods have been applied to address the issue of reference data error (section 4.3) and to produce spatially detailed depictions of map quality (section 7.4).

- *Reference data from a non-probability sample can potentially be used to supplement a probability sample for accuracy assessment.* Examples of such data are opportunistic windshield surveys or volunteered geographic information, collected in the locations convenient for data collectors. Stehman et al. (2018) describe approaches for incorporating such non-probability sample data sources within a design-based framework to improve precision of the estimates, but the improvement is generally not substantial.

2.3 Land cover, land cover change, land use

Land cover is defined as the observed (bio)-physical cover on the Earth's terrestrial surface. It includes vegetation and artificial features as well as bare rock, bare soil and inland water surfaces (Di Gregorio, 2005). *In situ* and satellite-based land observation efforts use land cover as one of the most obvious and detectable indicators of land surface characteristics and the associated human-induced or naturally occurring processes (Herold et al., 2009). Accurate land cover maps are required for understanding and mitigating climate change, monitoring of habitat and biodiversity loss, deforestation, land abandonment, conversion to agriculture, and urbanization (Radeloff et al., 2024; Wulder et al., 2018).

Definitions of land cover classes in a map should be **mutually exclusive**, and the legend **comprehensive** (or include an 'other' or 'unknown' class), in order to enable a transparent and unambiguous validation process (Wulder et al., 2018). If, for example, the set of classes is incomplete (not exhaustively defined), the accuracy of a classification may differ from the accuracy of a map generated from the classification (Foody, 2021); cases of an untrained class will typically be commissioned into the set of trained classes. In that respect, the Land Cover Meta Language (LCML) developed by FAO (its precursor being the well-known Land Cover Classification System (LCCS)) is recommended as the very first ISO standard (ISO 19144-2:2012) developed with the aim of providing a common reference structure for the comparison and integration of data for any generic land cover classification system (Di Gregorio and Leonardi, 2016). This tool can be used to create and describe land cover classes in a standardized and consistent way, thus facilitating the inter-comparison and validation of maps. It has been designed to be flexible enough to accommodate new land cover features and concepts in the future. Others pointed to potential limitations of classification systems such as LCCS and LCML (Câmara, 2020; Jansen et al., 2008).

For global applications, land cover class definitions should be broad enough to enable **consistent application across geographic domains** or include specific **subclasses reflecting regional differences**. The latter case will lead to a larger number of classes, which could pose accuracy assessment challenges related to larger number of sampling strata and a larger sample size to quantify per-class accuracy, as well as challenges related to obtaining reference data reflecting geographic variation in land class definitions.

When mapping land cover using satellite imagery, **the definition of land cover classes is related to the spatial resolution**. In coarse-resolution maps (resolution of 100+ m), it is often not possible to identify a single land cover class per pixel due to multiple land cover classes being mixed within a very large pixel (Townshend, 1992; Wulder et al., 2018). Such maps require higher resolution (e.g., 10-30 m) reference data

to validate sub-pixel proportions of land cover classes (e.g., Tsendbazar et al., 2018 and Figure A.1.2 in the Appendix).

With 10-30 m resolution maps, land cover class attribution to individual pixels is less ambiguous, as land cover generalization in the imagery at that scale has closer correspondence to the conventional landscape-scale land cover classes, such as 'forest', 'cropland', 'grassland', etc. The boundaries between classes are sharper, which allows use of the same resolution data for both mapping and validation (e.g., Landsat to validate Landsat-based maps) as long as the method of deriving reference sample labels is more accurate than the mapping method (e.g., visual sample interpretation vs. automated classification, Olofsson et al. (2014)). Often using the same resolution data is the only way of validating historic land cover maps, e.g., Landsat-derived maps before 2000, when sub-meter resolution data were not available. If higher resolution (<10m) data are available for all sample locations, it should be used as a primary source of reference data to validate sub-pixel land cover proportions in 10-30m resolution maps. However, it is often not the case, as the availability of public very high-resolution data (e.g., in Google Earth Pro or Bing Maps) varies among the geographic regions (see section 6.1) and in time (e.g., better coverage for recent years; no guarantee of getting reference imagery for all locations within the same year).

As we are entering the era of wall-to-wall very high-resolution land cover mapping enabled by the constellations of mini-satellites (e.g., PlanetScope, SkySat, BlackSky), individual features discernable in satellite imagery (e.g., tree crowns or even branches, separate buildings with lawns in between, etc.) are becoming finer than landscape-scale land cover definitions. This brings new challenges to land cover mapping and validation opposite of those at coarse spatial resolutions: now the question is, how to combine the land cover components such as 'bare soil' or 'vegetative cover' into meaningful land cover classes (e.g., tree cover and gaps in between forming the 'forest' class, or buildings, road, parking lots and city greenery forming the 'urban' class) or perhaps whether to map and monitor the composition of individual land cover classes (e.g., to detect changes in city structure). Novel deep learning classification methods that consider spatial context in addition to the spectral characteristics of the pixels, are enabling mapping of land cover using very high-resolution data (Feng and Li, 2020), but the accuracy assessment methods (particularly - aspects of response design, see Chapter 4) will need to be adapted to address the scale-related land cover definition challenges.

Regardless of the resolution of the satellite imagery used for mapping, land cover definitions may include Minimum Mapping Unit (MMU) or map spatial support unit, meaning a minimum area of a patch that is used to define a land cover class. For example, forest is often defined as a cluster of trees larger than X ha, which has implications for monitoring deforestation (Zalles et al., 2024). MMU can be used in both pixel- and polygon-based (also called object-based) land cover maps, and various land cover

classes on the same map might have different MMUs. Inclusion of a MMU into a land cover definition complicates accuracy assessment, as it needs to be considered while identifying a reference condition of a sample unit (Radoux and Bogaert, 2017). See introduction of Chapter 4 for further discussion on the relationship between spatial support unit of accuracy assessment and of the map, and how it relates to sampling unit and assessment unit.

Land cover change includes the conversion from one land cover category to another (Riebsame et al., 1994) and the modification, or subtle within-class change, that affects the character of the land cover without changing its overall classification (Coppin et al., 2004). From the temporal perspective, land cover change can be **ephemeral**, **interannual** or **semi-permanent/permanent** (Strahler et al., 2006). Ephemeral changes are short-term changes in cover, such as floods or seasonal burning in a savanna setting, which do not permanently alter the dominant vegetation cover distribution of the landscape. Interannual changes are variations in land cover largely due to long-term climatic variability, such as change in the annual extent of grasslands in the Sahel or reduction of woodland canopy cover for an area experiencing long-term drought. Semi-permanent/permanent changes include, for instance, new construction of impervious surfaces, deforestation events, or the expansion of agricultural lands. Land cover modifications, as compared to land cover conversions, are a form of semi-permanent/permanent change within a given land cover category. This is a more subtle form of change and includes examples such as rangeland degradation due to overgrazing, forest thinning due to selective logging and agricultural intensification.

Land cover change maps typically characterize land cover conversions or modifications, with the former being less ambiguous from a validation standpoint. Ephemeral changes have not been traditionally a target of land cover change mapping, but the emerging near real-time land cover mapping (Brown et al., 2022) presents these ephemeral changes to users as land cover changes without the explanation of their temporary nature or proper temporal validation, which is potentially misleading. The more temporary the land cover change that is being mapped, the harder it is to derive reference data capturing this change (e.g., contemporaneous field data or high-resolution satellite imagery). The inherent challenges of land cover change map validation are discussed in the following section 2.5.

Land use characterizes the arrangements, activities and inputs people have undertaken on a certain land cover type to produce, change or maintain it. Ideally, land cover and land use should be dealt with separately, but this is most often not the case. Because of the implicit or explicit role of humans in land use characterization, land cover and land use are not clearly distinguished and land cover maps tend to include both concepts (Chazdon et al., 2016; Comber et al., 2008). For example, in the case of non-vegetated surfaces, the spectral and temporal signal observed in the satellite imagery

may be used within classification rules to separate barren ground, snow or ice, and different human-created land covers (types of agriculture, such as classifying lands by crop rotations). Yet, as human activity and land uses cannot be sensed directly, some types of land use might be recognized with more difficulty, or even selectively omitted from maps (e.g., agricultural pastureland being misclassified as naturalized grassland habitats) (Burnicki, 2011; Steele et al., 1998).

During map accuracy assessment and intercomparison, **the difference between land cover and land use-based class definitions need to be accounted for** and reflected in the response design. A common example is forest (Zalles et al., 2024) defined as a land use (FAO, 2020), when temporarily deforested areas that remain under forest land use (e.g., clearcuts in forestry operations) are considered unchanged forest, vs. defined as the presence or absence of tree cover (Hansen et al., 2013), when all permanent changes in tree cover are mapped as forest disturbances. Mapped classes defined based on land cover are possible to validate using remote sensing imagery only, whereas accuracy assessment of land use-based maps requires additional information on the intended use, which cannot be derived directly from the imagery.

2.4 Categorical maps vs. continuous fields

Image classification is the process whereby individual pixels and/or groups of pixels within an image are categorized and labeled according to user-set or algorithm-derived rules. The primary units for characterizing land cover are categories (e.g., pixel classified as forest or as water) or continuous variables (e.g., % forest cover in a pixel).

Categorical maps divide land cover into discrete categories, or classes. Defined as the thematic resolution, categorizations range from being coarse (e.g., just forest) or fine (e.g., separating coniferous from deciduous vegetation). For example, in the United States National Land Cover Dataset, developed land is further differentiated into classes of land use by ‘open space’, ‘low intensity’, ‘medium intensity’ and ‘high intensity’ development (Dewitz, 2021). CORINE Land Cover, which is produced for the European Union as a part of the Copernicus Land Monitoring Service, provides a 3-level hierarchical land cover legend including, for example, the ‘artificial areas’ 1st level class, ‘urban fabric’ 2nd level class, and ‘continuous urban fabric’ 3rd level class (Kosztra et al., 2019).

Comparatively, **continuous fields** in land cover maps are used with the aim of better expressing the gradual transitions between landscapes and representing highly fragmented landscape mosaics. Continuous land cover maps are those where pixel values are not categorized but rather exist across a range of values (e.g., % tree cover in a pixel) (Cushman et al., 2010; Hansen et al., 2013). In a continuous field, the assertion is that each patch of similar cover, whether that be specific (old-growth Pacific conifers) or general (natural land), is not homogeneous at any thematic resolution, but rather, has

heterogeneous traits (Gao et al., 2019; Lausch et al., 2015). Using this model, pixels have the ability for either multiple (e.g., trees *and* coniferous) or partial (e.g., percent composition of that pixel composed by trees) class membership. One of the most widely used global continuous fields multi-class land cover maps available for the last 20+ years is MODIS Vegetation Continuous Fields (VCF), which is an annual 250 m resolution global map of surface vegetation cover as gradations of three components: percent tree cover, percent non-tree vegetation cover, and percent bare ground (DiMiceli et al., 2022).

While continuous fields are more specific than categorical maps, generalizability of landscape remains a valuable tool for conservation, land management, and planning applications, which is often difficult to discern from continuous land cover maps alone. Thus, a methodological pluralism rightfully exists, as the relative benefits to understanding land cover composition and configuration generally, is well complemented by continuous fields when investigating specific processes occurring in terrestrial and aquatic landscapes. Nevertheless, it is important to keep in mind that continuous fields are more difficult to validate, and **these guidelines focus primarily on the categorical classes**. Strahler et al. (2006) describe alternative metrics of accuracy for soft and fuzzy classifications (section 2.5), which result in maps of the fractions of thematic classes in a single pixel. There are various approaches to evaluating the accuracy of continuous fields, including modifications of the standard confusion matrix approach, commonly used to evaluate categorical maps (Binaghi et al., 1999; Foody, 1996; Woodcock and Gopal, 2000) as well as other metrics to describe accuracy (Pengra et al., 2015; Riemann et al., 2010).

One difficulty when assessing the accuracy of maps of continuous fields is deriving reference values of sub-pixel fractions of land cover classes (see section 4.1). Increasing availability of high-resolution satellite (see section 6.1), airborne lidar (section 6.2), and UAV (section 6.3) data facilitates this process. For example, reference sub-pixel fractions of tree cover within a Landsat pixel can be derived from circa 1 m resolution optical data (Potapov et al., 2017). The main challenge for a statistically rigorous validation is then obtaining these high-resolution reference data for the entirety of the probability sample. New constellations of high-resolution satellites (e.g., Planet's Dove) providing frequent global data coverage are therefore a valuable asset for the land cover mapping community, if free data access is provided to the scientific community through programs like Norway's International Climate and Forests Initiative (NICFI) satellite program (NICFI Data Program, 2021).

2.5 Land cover change maps

In contrast to the single snapshots in time provided by a static land cover map, land cover change maps rely on successive revisits of Earth observing satellites over the same area. In the past, the focus has been on the generation of maps for baseline years, e.g., GLC-2000 for the year 2000 (Fritz et al., 2003), GlobCover for the years 2005 and 2009 (Defourny et al., 2009) and the MODIS collection of annual land cover maps from 2000 to 2018 (Sulla-Menashe et al., 2019), which were not intended for change detection. However, there is a need for dynamic information to monitor changes over time, thus the emergence of new products that address this need. Land cover change detection is currently addressed by the community through **class-specific approaches** or **generic (multi-class) approaches**. The former is sometimes done by creating annual maps of a land cover class and then deriving land cover changes by comparing these annual maps (e.g., Pickens et al., 2020) or sometimes by directly mapping the change class (e.g., Hansen et al., 2013). The mapping of multi-class land cover change is usually performed by creating annual (or 5-year, or decadal) land cover maps, and then deriving the change matrix (containing all from - to land cover change categories) by comparison of individual land cover maps for two or more years or time periods. Ideally, the land cover change mapping should be addressed independently from the land cover mapping to avoid the impact of classification error propagation.

The forest loss and gain maps developed using Landsat time series data provide a detailed record to track afforestation, reforestation and deforestation globally (Hansen et al., 2013). Focusing on the pantropical scale, the Tropical Moist Forest dataset also derived from the Landsat archive allows characterizing the forest extent and changes (including forest degradation) from 1990 (Vancutsem et al., 2021). Similarly, the Landsat-based time series of water can be used to monitor water extent and change (Pekel et al., 2016), and is now maintained within an online application to support the United Nations (UN) Sustainable Development Goal indicator 6.6.1 (https://global-surface-water.appspot.com/). Cropland extent and change have also been addressed through the production of a Landsat-derived spatiotemporally-consistent dataset from 2003 to 2019 (Potapov et al., 2022b).

In addition to targeted monitoring of land cover change (e.g., forest loss), multi-temporal or annual land cover products are used to assess and/or estimate land change areas. The ESA CCI and EU C3S initiatives delivered the longest time series (1992-2020) of global land cover maps at 300 m spatial resolution for use in global climate models, mapping the land cover change component independently from the stable one thus avoiding propagation of classification errors (Defourny et al., 2017; ECMWF, 2020a, 2020b). More recently, the EU Copernicus Land Monitoring Service released another set

of global land cover maps over a shorter period (2015-2019) but at 100 m spatial resolution (Buchhorn et al., 2020a).

Whatever the approach, to interpret land cover maps and land cover change as part of climate change mitigation and adaptation, concern needs to be given to land cover change map quality and accuracy (Foody, 2002). It is **critical to assess the accuracy of land cover change detection and estimate the change area from a reference sample** of higher-quality land cover change information. In particular, caution needs to be exercised when monitoring land cover change based on post-classification comparison of annual or multi-temporal maps, since differencing maps with misclassification errors (e.g., 20% error in each map) leads to the erroneous detection of land cover change. Hence, statistically rigorous and transparent approaches for assessing the accuracy of land change monitoring are required (Olofsson et al., 2014).

Land cover changes are typically small in extent, compared to static land cover classes, therefore using **sampling designs that target land cover change classes through stratification** is especially important for improving the precision of the estimates (section 3.4). Omission errors in the maps used for stratification might have a significant impact on precision of area estimates for these small land cover change classes, which is somewhat mitigated by adjusting the sampling design, e.g., splitting large sampling strata into sub-strata, targeting areas of potential omission errors (Olofsson et al., 2020). Stehman and Foody (2019) discuss other challenges of sampling design for estimating the accuracy of land cover change maps related to the **type of land cover change (cyclical or unidirectional)** and to **whether the accuracy assessment needs to be repeated** in the future. Two sampling designs are discussed: a single, permanent set of sample units observed each year, and a different sample for each annual change estimate, with both approaches having their advantages for different estimation objectives. Section 7.1, based on Tsendbazar et al. (2021), proposes a framework for operational validation updates with additional strata targeting land cover change areas being added to the original sampling design.

One of the challenges when validating land cover change relates to **reference data availability**. For historic land cover change assessments, the data being used for mapping (e.g., Landsat archive) are often the only source of reference data. Some properties of land cover change, e.g., land use after disturbance, can be validated in the field after the land cover change has occurred (Krylov et al., 2018), but usually there is no way of predicting the areas of land cover change to coordinate field data collection as land cover changes are happening. High-resolution data (<10 m) had not been a reliable source of reference data for validating land cover change maps until the emergence of constellations of high-resolution satellites with frequent global acquisitions (e.g., 5 m RapidEye constellation in 2008-2020 with ~6 day nadir revisit time, and 3-5 m Planet

Dove constellation providing PlanetScope imagery since 2014, currently with near daily global coverage).

Another challenge of land cover change map validation is the **large number of transition classes to be validated**. These transition classes may represent very small areas and consequently yield lower precision of the accuracy and area estimates due to the small per class sample size. This is especially relevant for validating the land cover change derived from the comparison of multi-class land cover maps for different dates. This could be solved by reporting accuracy and area estimates for the aggregated land cover transition classes, e.g., 'forest to any class' (i.e. forest loss) or 'any class to forest' (i.e. forest gain) (Stehman and Foody, 2019). When assessing the accuracy of mapping land cover change trajectories for multiple dates, standard accuracy metrics might be less useful. In these cases, Stehman and Foody (2019) suggest to estimate the **summary land cover transition metrics** proposed by Pontius et al. (2017) from the reference sample data and to compare them with the same metrics derived from the maps. The proposed land cover transition metrics are: 1) the number of land cover change event incidents during the study interval; 2) the number of different land cover classes that the pixel belongs to over all points in time; and 3) the flow matrices expressing transitions of one land cover class to a different class between two points in time.

2.6 Map accuracy assessment vs. other map and model quality assessment methods

Map accuracy assessment

Map accuracy assessment is defined in the current protocol as the evaluation of the quality of the map based on comparison of the map labels with the independently derived reference labels for a probability sample of locations. Map accuracy assessment is also referred to as 'validation' (Justice et al., 2000), and both terms are used interchangeably in this document. Following Stehman and Foody (2019), the main focus is on the correct use of design-based inference for the estimation of **overall** (characterizing the entire set of land cover classes, e.g., overall accuracy) and **class-specific** (characterizing a specific land cover class, e.g., user's and producer's accuracy) **accuracy metrics**. Overall and class-specific accuracy metrics can be computed for the entire map, or for subregions of the study area (e.g., individual countries or regions within the country), which is a recommended approach to assess regional variation of map accuracy (see sections 2.7 and 5.1).

We primarily focus on validation of single maps, but **comparative accuracy assessment of multiple maps** is also recommended to be performed in a design-based framework based on an independent probability sample (section 2.8), and is therefore not

fundamentally different from a stand-alone map accuracy assessment, despite the specifics of sampling and response design (e.g., using stratification targeting areas of map disagreement and defining rules of harmonizing map and reference classification legends).

Estimation of the area of the target land cover class is also included in this protocol on map accuracy assessment (section 5.2), as sample-based area estimates provide insights into the possible utility of the map for pixel-counting, and area estimates can be derived from the same reference sample used for map validation with little additional effort. However, if area estimation is the primary objective of the study, sampling and response design specifics might be different, compared with an assessment primarily focused on map accuracy assessment (Jonckheere et al., 2024). At the same time, a sampling design selected to increase the precision of accuracy estimates will still be valid to estimate the area of land cover classes, and vice versa, as the choice of the sampling design will not introduce any biases into the estimation, as long as it is a correctly implemented probability sampling design. Area estimation utilizing design-based inference also includes **model-assisted estimators**, employing a model to obtain the benefit of enhanced precision of the estimates (section 5.2).

Other map quality assessment methods

Local map quality assessment methods employ models to produce quality metrics for each map unit (e.g., pixel). Such local quality metrics are often desired by map users, e.g., when land cover maps are used as inputs to biosphere or climate models. These methods include **interpolation approaches relying on sample data** (not necessarily on a probability sample) and **map quality metrics produced from classification outputs**, see section 7.4 for an overview of these methods.

Systematic map quality control (section 2.10) could be performed during the map production to identify regions of poor model performance and to improve mapping results. **Intercomparison of maps** (section 2.9) could help verify map quality even if it does not follow a formal comparative accuracy assessment procedure. All these methods provide useful insights into map quality but should not replace the design-based accuracy assessment.

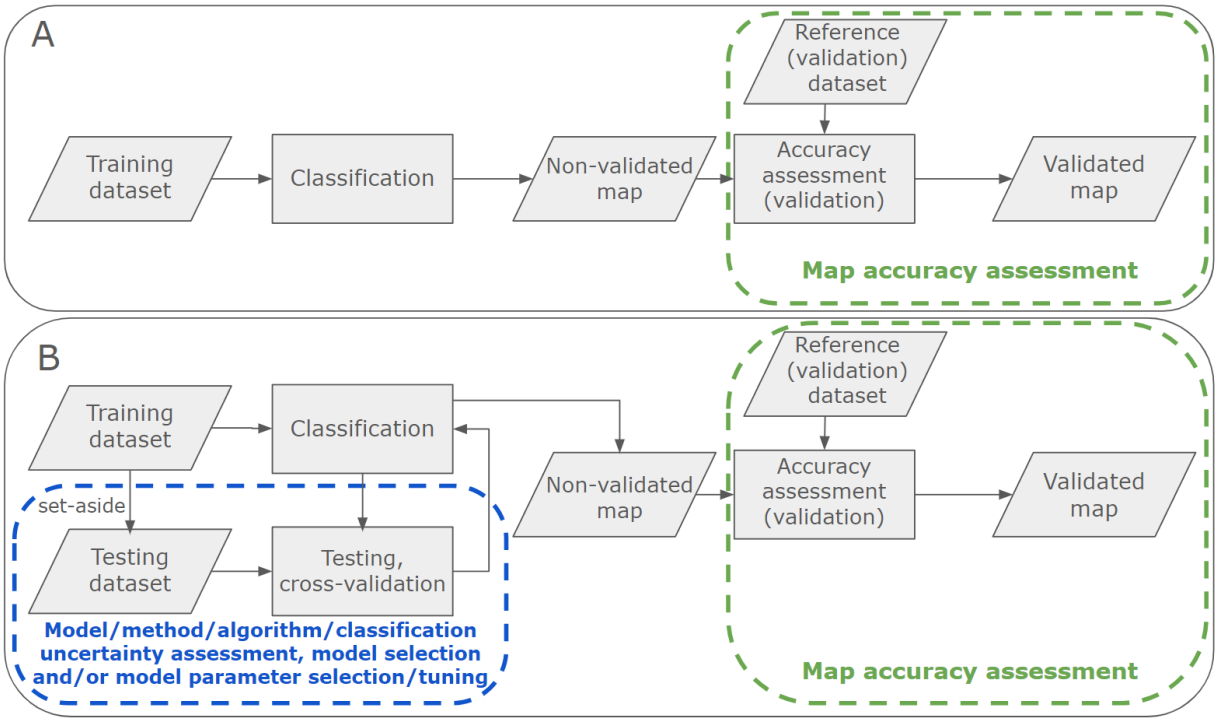

Figure 2.6. Workflow of map accuracy assessment (A) and map accuracy assessment preceded by model/method/algorithm/classification uncertainty assessment, model selection and/or model parameter selection/tuning (B). Map accuracy assessment should be based on a probability reference sample, independent from the map training dataset. The map training dataset can be subdivided into training and testing datasets in the workflow presented in (B), but this testing set-aside subset of the training dataset should not be used as a substitute for an independent reference dataset for map accuracy assessment, unless both the training dataset and its subset (testing set-aside) are derived via probability sampling. Even if this condition is met, using the set-aside of the training dataset for assessing map accuracy is not recommended (see Table 4.1), because errors in the reference data will be correlated with the errors in training data and in the map, which will likely result in overestimation of map accuracy. Figure by Alexandra Tyukavina and Anna Komarova.

Model/method/algorithm/classification uncertainty assessment

Model/method/algorithm/classification uncertainty assessment aims to evaluate the performance of a classification model to estimate how well this model will generalize to data not used to train the model. Such evaluation, also referred to as 'testing', supports the choice between model types, the tuning of parameters and features, etc. It can also be used to identify poor representation of training data for some classes and assess the classification performance for marginal or less frequent classes (Figure 2.6). Testing is usually performed on a subset of the training dataset that is set aside so that the model can be trained and tested on different data. Model uncertainty assessment or testing should be based on the training dataset only, and not on the validation (reference) dataset, which must remain fully independent for map accuracy

assessment. Testing data can be used for cross-validation or estimating the out-of-bag error (Colditz et al., 2011; Friedl et al., 1999; Li et al., 2014; Pouliot et al., 2009; Praveen et al., 2023; Radoux et al., 2014). This procedure has been widely used due to its speed, simplicity, and flexibility. Nevertheless, it is important to clarify that **this is NOT a map (product) accuracy assessment as defined in these guidelines**, but an evaluation of the classification model performance. The conditions under which the testing dataset (subset of training data) could be used as a substitute for an independent reference dataset for map accuracy assessment are discussed in the caption of Figure 2.6.

2.7 Accuracy metrics and area estimates

In principle, the estimation of map accuracy and the area covered by a class within the region mapped are straightforward tasks. Accuracy is fundamentally a measure of quality that is focused on the degree of thematic error contained in a map. The area of a class is often simply interpreted as being the total spatial extent of the class within the region mapped. However, both accuracy and area may be estimated and characterized in a variety of ways and care is needed to ensure that a rigorous and appropriate methodology is used to avoid bias and to allow estimation of uncertainty of the accuracy and area estimates.

Accuracy metrics

The presence of map errors impacts both accuracy assessment and area estimation. A simple summary of the correct and erroneous allocations in a map can be obtained by cross-tabulating the observed map label with the corresponding actual label observed in a ground reference dataset for a sample of cases drawn from the region mapped; the specific nature of the sample (i.e. sampling design) is very important and is discussed later in Chapter 3. This cross-tabulation is typically referred to as an **error or confusion matrix** and is the basis of recommended methods of accuracy assessment and area estimation (Olofsson et al., 2014; Strahler et al., 2006).

The main diagonal elements of the confusion matrix correspond to the cases (e.g., pixels) that have been correctly allocated in the classification used to generate the map (i.e. these cases have the same class label in the map as is observed on the ground). The off-diagonal elements of the matrix illustrate errors (cases for which the map and actual class label differ). Various accuracy metrics can be generated from the confusion matrix (Liu et al., 2007). The most widely used map accuracy metric is **overall accuracy**, expressed as the proportion (or percentage) of correctly mapped area (Olofsson et al., 2014; Trodd, 1995). Such a metric gives an overall guide to the quality of the entire map (single measure applied to the entire set of classes) and because of this property is sometimes described as being a global metric, even though it can be computed for subregions or subdomains. To avoid confusion with the geographic scale at which these

metrics are computed, we will refer to the accuracy metrics characterizing the entire set of classes as overall accuracy metrics. Some overall metrics have been recently recognized as inappropriate for accuracy assessment, for example, the kappa coefficient of agreement. Kappa has been widely used in the past (Congalton and Green, 2019) and can sometimes be generated in popular image analysis software, but it makes an unnecessary correction for chance agreement, and is highly correlated with overall accuracy, and therefore is not a recommended metric (Foody, 2020; Olofsson et al., 2014; Pontius and Millones, 2011). However, as a range of metrics exists and various metrics may meet specific needs of the mapping community, **it is recommended that the confusion matrix be provided** as part of the accuracy statement allowing map users to calculate those metrics of interest (Olofsson et al., 2014; Stehman and Foody, 2019). In addition, **confidence limits should be reported along with estimated accuracy metrics** (Olofsson et al., 2013).

In practically all applications, there is interest in the accuracy with which individual land cover classes have been classified. A range of **metrics of per-class (class-specific) accuracy** can be calculated from a confusion matrix (Figure 2.7). A classification error occurs when a case (map unit) is mis-labeled, as there are many sources of error in a land cover map (Foody, 2002). Note, however, that a single mis-labeled case is associated with two different types of error. First, the case may belong to class X (ground condition) but be erroneously mapped as class Y; an error of omission from class X. Second, the same case has been erroneously mapped as class Y from class X; a commission error into class Y. Thus two different perspectives on map accuracy can be observed, commonly referred to as **user's accuracy (complement of commission error)** and **producer's accuracy (complement of omission error)** (Figure 2.7, section 5.1). Different expressions for these metrics can be observed in other communities (e.g., machine learning) as for instance, producer's accuracy may be called 'sensitivity' or 'recall', and user's accuracy called 'precision' (see Table 5.1.2). User's accuracy is calculated from the ratio of correctly mapped area of a class (cells of the confusion matrix for which the map and the reference labels agree on the presence of the target class) to the total mapped area of that class. Producer's accuracy is calculated from the ratio of correctly mapped area of a class to the total area of that class determined from the reference classification (sum of cells of the estimated confusion matrix corresponding to the target land cover reference class). It is also possible to summarize user's and producer's accuracies in a single value, the best known being the F1 score (see section 5.1). F1 score, and other unique metrics (Liu et al., 2007), have the advantage to facilitate the comparison between classes and maps but they are based on the premise that commission and omission errors have the same importance while the relative importance of the different types of accuracy may differ depending on the intended map uses and the user community needs. **Therefore, it is recommended to always provide, as a minimum, the user's and producer's accuracy metrics.**

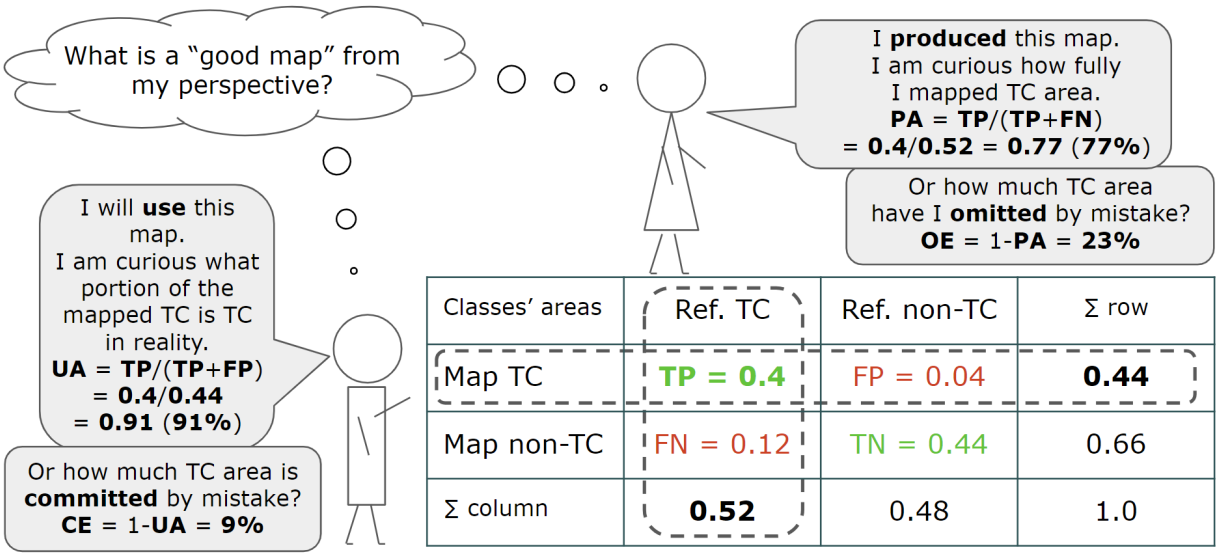

Figure 2.7 Conceptual diagram illustrating calculation of per-class (class-specific) accuracy metrics from a binary (two-class) confusion matrix, for a Target land cover Class (TC). The diagram also presents a mnemonic rule for memorizing the difference between user's (UA) and producer's (PA) accuracy, presenting objectives of a map producer and potential users on evaluating map quality. TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative. CE is commission error rate; OE is omission error rate. Error matrix is presented in terms of area proportions as recommended by the current protocol (see Chapter 5 for more details). Diagram by Anna Komarova and Alexandra Tyukavina.

Providing the confusion matrix is also helpful in applications when there is a desire to collapse classes for application-specific reasons. This situation can be aided by using a hierarchical classification scheme. For example, a map may show coniferous and deciduous forests, but a map user might want a single class of forest. This simply requires the cases of the relevant classes to be relabeled and revision of the confusion matrix to show the new, smaller, set of classes (Strahler et al., 2006).

Please note that both **the overall and the class-specific accuracy metrics can be computed without explicitly constructing a confusion matrix**. These are direct sample-based estimators, based on comparing map and reference for each pixel, and then doing weighted summation over the strata for stratified sampling designs (e.g., Stehman, 2014). These estimators might be more convenient if the sampling design is complex (e.g., multiple strata not corresponding to map classes or sampling with unequal inclusion probabilities). Section 5.1 refers to papers containing accuracy estimators appropriate for various sampling designs.

Overall and class-specific accuracy metrics described above only provide a summary for the entire mapped region. However, **accuracy is known to vary in space**. This spatial variation in accuracy may arise as a function of, for example, spatial variations

in the nature of image acquisition (e.g., due to persistent cloud cover or variations in angular viewing geometry) or of the landscape mosaic and environmental variables. Consequently, map accuracy is not always well characterized by a single accuracy metric for the entire map. Additionally, there is sometimes interest in specific locations (sub-regions) within the mapped region. The recommended approach is to simply **focus a standard accuracy assessment on the sub-region of interest** to identify the subset of the larger-scale sample falling within a sub-region of interest and increasing sampling density in the sub-region to achieve desired precision of the estimates (see section 3.5 for the overview of the methods for calculating required sample size depending on desired precision of the estimates). This method would then produce a single accuracy estimate for the entire sub-region of interest (or estimates of user's and producer's accuracy for each land cover class within a sub-region).

An alternative is to characterize the variation of classification quality in space. A variety of approaches may be used to illustrate the **local variation in accuracy** (at the scale of individual map units) often based on interpolating between the sites of ground reference data collection or spatially constrained analyses (Comber et al., 2017, 2012; Comber and Tsutsumida, 2023; Ebrahimy et al., 2021; Foody, 2005). Another approach is to generate maps of the uncertainty of class allocation (class assignment for each map unit). Although this is a different measure of map quality than accuracy, the spatial variation in class allocation uncertainty can provide spatially explicit information on map quality to enhance an accuracy assessment (Foody et al., 1992; Steele et al., 1998). See section 7.4 for an overview of methods characterizing local variation in accuracy via the interpolation of the reference sample and other local map quality assessment methods.

Map accuracy **might also vary in time**. For multi-year land cover maps, standard accuracy assessment with a single sampling design could be used to assess accuracy of the map for individual years or epochs. In this case map epochs could be used as sampling strata. Existing sampling design can also be modified for operational map updates (see section 7.1), with additional sample units targeting the new areas of change. In this case a protocol for revisiting existing sample units (or their subset) needs to be established.

Accuracy assessment is very much **a developing subject** and a topic of ongoing research. A variety of issues require consideration if rigorous estimates are to be obtained and reported (Stehman and Foody, 2019). For instance, recent research has included work on methods to express and present accuracy information (Meyer and Pebesma, 2022) and explored topics such as the potential for model-based approaches to accuracy assessment and area estimation (Foody, 2012; McRoberts, 2010; McRoberts et al., 2022; Steele et al., 2003). In addition, the basic methods may be adapted to account for particular circumstances. For example, the standard approach to accuracy assessment weighs all errors equally but there may be times when some errors are more impactful

than others and a weighting approach is desired (De Bruin et al., 2001; Stehman, 1999). Also, the approach may need to be adopted for non-standard classifications such as when the map uses a soft or fuzzy classification or continuous fields of land cover variables rather than a categorical map (see section 2.4). With soft classifications, a case can have multiple and partial class membership, and the conventional approach to accuracy assessment based on a confusion matrix for hard classifications (in which each case is associated with just one class in the map and one class in the ground reference data), is inappropriate. A range of approaches for the evaluation of soft classifications, including adaptations of the basic confusion matrix approach, exist (Binaghi et al., 1999; Foody, 1996; Woodcock and Gopal, 2000). This type of accuracy assessment is of particular relevance to maps of, for example, continuous fields or even just situations in which mixed pixels may be common.

Thus, in terms of **accuracy metrics, the key recommendations** are to summarize **overall accuracy** using the correctly mapped area and express **class-specific accuracy from both the user's and producer's perspectives,** reporting both types of accuracy metrics **with their estimated confidence limits.** Provision of the **confusion matrix** with these summary statistics **allows the calculation of other metrics if desired.** The confusion matrix should be expressed in area proportions instead of sample counts if the sampling design is not equal probability (e.g., sample unit inclusion probabilities differ by strata) (Olofsson et al., 2014; Stehman and Foody, 2019). Matrix processing operations (e.g., matrix normalization) can have detrimental effects and should not be used (Stehman, 2004).

Area estimates

Area estimation is often achieved by summing the extent of all patches in the map that have been allocated to the class of interest. Thus, for example, with a standard per-pixel classification, the area of a class can be estimated by counting all of the pixels in the map that have been allocated to the class and multiplying this value by the pixel size (typically expressed in units such as m^2). The area would typically then be reported in km^2 or as a proportion (or percentage) of the mapped area. This estimate of area is the area of the class as shown in the map and is often referred to as a **map-based area estimate** or **map pixel counting**. It is, however, a naïve estimate and not necessarily a good indicator of the actual area of the class because it **is biased by mis-classification errors in the map**. Instead, the **area** of a target land cover class **should be estimated from the sample data of reference conditions (reference classification)**, which should be obtained independently from the map, and using a more accurate labeling method.

The area estimation process needs to account for the effect of mis-classification bias in the map as it can be a source of misestimation of class area. As a simple

illustration, if a class suffered no omissions but commissioned many cases from other classes, then its area would be inflated by use of a basic pixel counting approach. Fortunately, the confusion matrix used to generate an accuracy statement also contains the information needed to adjust area estimates for mis-classification bias (Olofsson et al., 2014, 2013). However, the 'error-adjusted' (Olofsson et al., 2013) or 'bias-adjusted' (Stehman, 2013) area estimate terminology is somewhat misleading, as it over-emphasizes the role of the map in producing sample-based estimates (Stehman and Foody, 2019): these **area estimates can be produced from the reference labels directly** (or from the cells of the estimated confusion matrix corresponding to the reference class), without the need of 'adjusting' the map-based estimates. The role of the map in sample-based area estimation is to reduce standard errors via stratification (see section 3.4 for more information about the role of stratification). Section 5.2 refers to publications providing direct sample-based area estimators for various sampling designs which could be used to produce area estimates even when maps of the target land cover class are not available (e.g., simple random or systematic sampling or stratified sampling using desired reporting regions rather than target map classes as strata).

A confidence interval should also be estimated to quantify the uncertainty associated with the sample-based estimate of class area (Olofsson et al., 2013). Estimators of variance for area estimates, appropriate for each sampling design, are typically provided along with the area estimators, and enable constructing confidence intervals. Selection of an appropriate confidence interval estimation method is an ongoing area of research (Stehman and Xing, 2022). Another recent focus of research is estimation of area change (e.g., deforestation) (Olofsson et al., 2020) and the potential of model-based methods for area estimation (Foody, 2012; McRoberts, 2010, 2006).

2.8 Comparative map accuracy assessment

Comparative map accuracy assessment (validation) methods follow the procedure of a sample-based accuracy assessment using design-based inference. The focus is, however, on **validating multiple maps using the same reference dataset and harmonized land cover class definitions**. The motivation for performing comparative validation is that comparing reported map accuracies produced for each map separately may not be straightforward due to the differences in the reference datasets and class definitions. Comparative validation (also referred to as 'benchmarking') is usually more challenging than validating a stand-alone map using a reference dataset created specifically for the validation of that map with matching land cover definitions. Stehman and Foody (2019, section 4.6) discuss statistical methods of testing whether the accuracies of two or more datasets estimated from the same reference sample are statistically significantly different or equivalent.

Comparative validation using **stratification, targeting the areas of disagreement between the maps,** and using the same reference dataset, allows highlighting the respective advantages or drawbacks of the different products. This concept is illustrated in Lamarche et al. (2017) who compared several global surface water products using a stratification defined to distinguish between high confidence in correctly mapping the land class (stratum 1 - 25% of the sample), high confidence in correctly mapping the water class (stratum 2 - 25% of the sample), and error-prone areas mainly corresponding to shorelines, lakes, and river banks (stratum 3 - 50% of the sample). The overall accuracy computed using all sample units was very high among all products, between 98% and 100%. Considering only the sample units targeting error-prone areas, all products yielded substantially lower accuracy numbers, and the differences between products became noticeable, with overall accuracies in the error-prone stratum ranging between 74% and 89% (see more details in section A.3 of the Appendix). In regard to this comparative validation example, comparing the overall accuracies of multiple maps will often fail to distinguish between nearly similar products, particularly in the case of strongly imbalanced binary maps with rare target classes (e.g., water bodies, deforestation). **Class-specific accuracy metrics** (see section 5.1), such as user's and producer's accuracy, **should be used to compare the quality of such rare classes from different land cover maps** instead of basing the comparison on overall accuracy.

Tsendbazar et al. (2016) demonstrated the use of an existing reference dataset, created for validation of the Globcover-2005 map, to perform a comparative validation of three maps of 300-500 m resolution for the year 2005. To do this comparison, they reinterpreted the land cover legend of the reference dataset into the different map legends. With the acceleration of the production of global land cover products over the past decade, there is a **need to collect validation datasets aimed to facilitate multi-purpose assessments** in order to save time and effort on data collection. The general principles for creating standardized reference datasets and associated challenges are discussed in section 7.3. The first multi-purpose land cover validation dataset was developed for Africa in the framework of the Copernicus Global Land Service (CGLS-LC100) (Tsendbazar et al., 2018), and then was expanded globally (see section A.1 of the Appendix for more details on this dataset). In Tsendbazar et al. (2018), the applicability of such a multi-purpose land cover validation dataset has been demonstrated in three different assessments focusing on (i) validating discrete and fractional land cover maps, (ii) map comparison, and (iii) user-oriented map assessments. The CGLS-LC100 reference dataset is also the basis of a comparative validation of 10 m global land cover maps (Xu et al., 2024).

2.9 Intercomparison of maps

Comparison between land cover products may also support the user's need to identify the 'best' available map (Herold et al., 2008): potential users of global land cover data question which map is the most useful for their purposes (spatial resolution, temporal update, thematic accuracy, etc.). Comparison between maps to identify their specific strengths and weaknesses is one option. Harmonization of multiple maps into a single improved global dataset is another option, as different maps can have higher accuracy of certain geographic areas, or higher thematic detail of certain land cover classes (See et al., 2015; Tsendbazar et al., 2015a). Map intercomparison and harmonization efforts are complicated by varying legends, which can be challenging to harmonize thematically (Vancutsem et al., 2012).

The first studies addressing this question compared global land cover maps to highlight their relative strengths (Fritz et al., 2011; Giri et al., 2005; McCallum et al., 2006). These studies analyzed spatial agreement among the maps, but they did not provide information on comparative accuracies of these maps. Fritz et al. (2011) and Herold et al. (2008) compared the accuracies of global land cover maps by harmonizing reported confusion matrices with different legends into confusion matrices with a common legend. As this harmonization only concerned confusion matrices reported by map producers using different reference data, it remained unclear how the accuracy of these maps would compare relative to the same reference dataset (see section 2.8 above for the example of comparative accuracy assessment studies employing the same reference dataset).

It is also important that these intercomparison exercises integrate the application domain in their protocol, i.e. how well each map suits different applications and meets the needs of various user groups. Conventional accuracy reporting from confusion matrices assumes that all confusion errors are equally important. However, confusion between certain classes may have more impact on applications of land cover maps than between other classes (DeFries and Los, 1999). Several studies accounted for such differences and calculated global land cover map accuracy for specific applications using weights derived from class similarities by parameters (DeFries and Los, 1999; Mayaux et al., 2006). More recent publications specifically assessed the strengths and weaknesses of maps for different applications with the twofold objective of (i) identifying priorities for improving the global land cover maps for those applications and (ii) helping users to select land cover maps most suitable for their applications and to understand their uncertainty (Tsendbazar et al., 2016, 2015b). Stehman and Foody (2019, section 4.6) further aspects of using accuracy data to compare maps, such as considering the width of confidence intervals and performing statistical tests to identify whether the accuracies of the maps being compared are statistically significantly different or equivalent.

2.10 Systematic map quality control

Map validation typically implies a statistical quantitative accuracy assessment, meaning comparison of the map with an independently derived reference sample. This is the type of assessment which has been discussed so far and which is the main focus of the current protocol (Chapters 3-5, section 2.7). However, it might be useful to complement rigorous statistical accuracy assessment with a systematic quality control protocol to help build confidence in the product.

Systematic quality control is defined as a process intended to meet two main objectives: eliminating macroscopic errors (i.e. errors that are visible for users but poorly detected by a statistical accuracy assessment, such as distinct clusters of commission errors) and increasing overall acceptance of the land cover product by users. Macroscopic errors reduce the user's overall confidence in the products, even if the quantitative accuracy is high (i.e., the total area of these errors is low relative to the area of mapped classes). The occurrence of such errors can be greatly reduced by a careful review of the products. Systematic quality control is also a way of assessing if the remotely sensed data have been correctly classified (i.e., if the errors are due to limitations of data quality rather than to poor classification procedures) and to investigate the influence of different variables (e.g., heterogeneity, class dominance) on the quality of the land cover map. Ideally, this 'early validation' step should be integrated into the classification procedure, to reduce the occurrence of macroscopic errors and improve the map. Systematic quality control is somewhat similar to the concept of 'active learning' in the classification of remote sensing imagery, where the map is visually evaluated after each classification iteration, and more training data can be added to improve map quality. Active learning, though, is a more general approach, and does not necessarily have to be performed systematically, as described below.

Systematic quality control was first documented by Mayaux et al. (2006) in the framework of the GLC-2000 validation. Qualitative validation was based on a systematic descriptive protocol, in which each cell of the map is visually compared with reference data and its accuracy documented in terms of type of error, landscape pattern and land cover composition. The grid size was adapted to the characteristics of the landscape, the map, and the reference data. Building on the GLC-2000 experience, such systematic quality control was also included in the ESA CCI and EU C3S validation protocols (Defourny et al., 2020).

3. Sampling design

Statistically rigorous accuracy assessment of land cover maps in a **design-based inference framework** (Stehman and Foody, 2019) is based on observations of reference conditions on the land surface at locations selected by **probability sampling** (see definitions in section 2.1). The same reference sample could be used to estimate the area of the target land cover class even if there is no land cover map available in the first place. This chapter covers various aspects of sampling design, which is "*the protocol by which the reference sample units are selected*" (Stehman and Czaplewski, 1998). We will discuss the elements of a sampling design: choice of the sampling unit (section 3.1), sampling frame (section 3.2), common probability sampling designs (section 3.3), the specifics of stratification (section 3.4) and sample size planning (section 3.5). Note that 'sampling design' is used here both in a broader sense including all the elements outlined above, and in a strict sense (section 3.3) meaning the protocol by which sampling units are selected.

Please note that "*because design-based inference does not assume independent observations, spatial correlation will not bias accuracy estimates, and the sampling design need not be chosen to avoid spatial correlation*" (Stehman, 2000). In other words, **in a design-based analysis and estimation framework, there is no need to space sample units apart to avoid spatial correlation** (Stehman and Foody, 2019). If ad hoc modifications of sampling design are made to avoid sampling of nearby units (e.g., randomly selected units are discarded or moved based on a proximity criterion), the inclusion probabilities of a modified sampling design can be very complicated and difficult to derive (Stehman and Foody, 2019), or even intractable. If modified inclusion probabilities are not accounted for, such sampling designs will no longer satisfy the criteria of a probability sampling design (see section 3.3) and the statistical rigor associated with use of design-based inference is forfeited.

To satisfy the principle of reproducibility (section 2.2), the **sampling design** employed **should be adequately described** when publishing accuracy assessment results (see Table 3.1 for the suggested sampling design metadata categories). When documenting the sampling design, it is recommended to (Stehman and Foody (2019), section 2.2, p.5):

1) describe the randomization implemented in the sample selection protocol;

2) specify the inclusion probabilities or the information needed to derive the inclusion probabilities [i.e., the area and number of units in the sampling region and strata sizes (if stratified sampling is implemented); selected sample size (and sample size allocation per stratum, if stratified)]*;*

3) if stratified sampling was implemented, describe how the strata were constructed, provide the proportion of area in each stratum, specify the sampling design

implemented within each stratum, and state the sample size allocated to each stratum along with the rationale for this allocation;

4) if cluster sampling was implemented, define the primary sampling unit (PSU) and the secondary sampling unit (SSU), state whether one-stage or two-stage sampling was implemented (in one-stage sampling all SSUs within each sampled PSU are observed, whereas in two-stage sampling a sample of SSUs is selected within each sampled PSU), and specify the sampling design implemented at each stage.

To satisfy the principle of transparency (section 2.2), **all deviations from the sampling design** (e.g., sample unit not visited in the field or replaced) **should be honestly reported** (Stehman and Foody, 2019).

Table 3.1. Suggested sampling design metadata to be reported to ensure transparency and reproducibility of map accuracy assessment.

Category		Description
Sampling unit	Required	The type of sampling unit used: points, pixels, fixed-sized polygons, unequal-sized polygons (see section 3.1).
Population and sampling frame	Required	Define target population and its size: total number of units in the population (*N*) and the size of these units if employing a list sampling frame, or the geographic extent of the sampling region and its area if employing an area sampling frame (see section 3.2).
Sample selection protocol	Required	Describe sample selection protocol (see section 3.3), including: 1) whether simple random or systematic selection protocol was implemented within the entire sampling region or within each stratum; 2) whether stratification was used; 3) whether clusters were used.
Stratification	Yes/No	If stratification is used (see section 3.4), include the following information: 1) how the strata were constructed; 2) area of each stratum; 3) sampling design implemented in each stratum (e.g., simple random or systematic sampling); 4) sample size allocated to each stratum and its justification (see section 3.5).

Category		Description
Cluster sampling	Yes/No	If cluster sampling was implemented, (see section 3.3) include the following information: 1) primary sampling unit (PSU) and secondary sampling unit (SSU); 2) one-stage or two-stage sampling; 3) sampling design implemented at each stage (see Figure 3.3.2).
Sample size planning	Recommended	It is recommended that required sample size is calculated depending on the estimation objectives (target precision of overall, user's, producer's accuracy or land cover class area) to avoid unnecessarily large sample sizes (see section 3.5). The following decisions related to sample size planning are recommended to be reported: 1) how the overall sample size was calculated; 2) (if using stratification) how the sample size was allocated among the strata; 3) (if working with multiple estimation objectives) how the decisions of prioritizing various estimation objectives were made; 4) any considerations leading to sample size modification/reduction, e.g., due to reference data collection costs.

3.1 Sampling unit

Sampling units are the entities that make up the sampling frame (Särndal et al. (1992), p.5). The literature does at times distinguish between population units and sampling units (e.g., Cochran, 1977, p. 6). While not always the same thing, the sampling frame is equivalent to the population in many situations (see section 3.2 for more details on the sampling frame). A **sample** is a subset of the units that comprise the population. The reference conditions are observed for the sample in the process of accuracy assessment. The main types of sampling units are **areal units**, such as **pixels or polygons**, which correspond to a finite area on the ground, and **points**, which have no areal extent (Stehman and Czaplewski, 1998). The sampling unit of the accuracy assessment does not have to match the spatial resolution or the primary mapping unit of the map being validated. A matching sampling unit (e.g., a pixel of a pixel-based map, or a polygon of a polygon-based map) is often selected for convenience of assigning reference labels and establishing correspondence between the reference labels and map values for each sampling unit. For example, if a point is used as a sampling unit, the

spatial resolutions of the map and the reference data will need to be accounted for when defining the response design protocol (see Chapter 4) to ensure that both the reference and map land cover values for each point correspond to the same area on the ground.

When **selecting the sampling unit** for the study, it is important to remember that "*no consensus exists on which sampling unit is best, and it is unlikely that any one sampling unit is optimal for all applications*" (Stehman and Czaplewski (1998), p. 333). Some of the considerations when choosing an appropriate sampling unit for a particular study are: **cost and time** of deriving the reference values for each sample unit, **sensitivity to geolocation error** (e.g., boundary pixels and map polygon boundaries), **ability to retain identity under map revisions** (e.g., map polygons may change with map versioning whereas fixed size grids do not change). Stehman and Wickham (2011) provide a comprehensive overview of pixels, square blocks of pixels, and polygons of varying sizes as sampling units, evaluating their respective strengths and weaknesses.

A **point sample** is a selection from a continuous (infinite) population not partitioned into discrete areal units via an area sampling frame (see section 3.2 below). The advantage of a point sample is its possibility of being re-used for accuracy assessment of multiple maps, since it is not linked to any particular pixel grid. When using the same point sample for assessing the accuracy of multiple maps, sample reference values will need to be revised depending on the land cover definitions and the spatial resolution of the maps being validated (see Chapter 4).

Pixels or **fixed-size polygon grids** (also referred to as 'pixel blocks', e.g., in Stehman and Wickham, 2011), e.g., 3x3 pixels, 10x10 km or 1 ha, are the common **areal sampling units**, reflecting the regular nature of grids typically used to store geographic data. Sampling pixels for validating pixel-based maps is a natural choice, as there is an unambiguous correspondence between the footprint of the sample and the map units. Blocks of pixels are typically used in one- or two-stage cluster sampling designs (see section 3.3 below), in which reference labels are assigned to all pixels within a block (one-stage cluster) or to a sample of pixels or points within a block (two-stage cluster). Larger pixel blocks can be beneficial when high-resolution imagery or ground reference data are acquired for each sampled unit, situations where blocks may achieve substantial reductions in cost of obtaining reference data.

A common misapplication of cluster sampling is to aggregate the pixel data within a block (cluster) and to compare map and reference class labels created at the block level rather than to compare the map and reference labels at the individual pixel level. An accuracy assessment based on aggregating data to the block level has an assessment unit (block) different from the units of the map (pixel). Accuracy results from such a block level assessment should not be used because they do not apply to the map (per pixel classification) distributed to users (Czaplewski, 2003). Therefore, **agreement** of

correspondence **between the map and the reference data** during accuracy assessment should be performed **at the scale of the map unit**, even if sampling is performed at the aggregated pixel block (cluster) level.

Grids of pixels and **fixed-sized polygons (blocks)** are defined independently of the thematic contents of the map being validated, and thus samples drawn from such grids retain validity over map revisions (Stehman and Czaplewski, 1998), unlike samples of polygons that are defined based on the land cover map class values. Note that when working with raster data in non-equal-area projections, regularly spaced grids (e.g., degree grid in geographic coordinates) are not necessarily composed of equal-area cells, and thus adjustments to the sampling design or analysis need to be made to account for the unequal sampling unit area. Fixed-sized polygons should ideally be aligned with the pixel grid of reference satellite data, to avoid ambiguity in reference sample labeling (Zalles et al., 2024). **Polygons of unequal size** are usually defined based on each polygon having homogeneous land cover classes either in the map being validated (**'map polygons'**) or in the reference data (**'reference polygons'**, Olofsson et al. (2014)). Polygon-based validation is sometimes referred to as **'object-based validation**' as opposed to 'pixel-based validation', which refers to the fact that polygon-based validation methods apply not only to the conventional wall-to-wall mapping results, but also to detection and delineation of separate objects in high-resolution imagery, such as individual trees, cars, buildings, crop fields (Radoux and Bogaert, 2017). Radoux and Bogaert (2017) summarize the good practices of validation using polygons of varying size, including specific accuracy estimators that account for the area of sampled polygons (Radoux et al., 2011).

3.2 Sampling frame

The **sampling frame** defines the target population from which the sample is selected. The two main types of sampling frames are **'list frames'** and **'area frames'** (Stehman and Czaplewski, 1998).

A **list frame** is simply a list of all possible sampling units (i.e., the population) in the target region, and therefore is suitable **only for sampling of areal units that comprise a population of finite size** (pixels, blocks of pixels, polygons) as opposed to an infinite number of points in a continuous population. When sampling from a list using simple random or systematic sampling, each unit has **equal probability** of being selected, and thus for populations of units with unequal area, the area of each individual unit needs to be accounted for when estimating the accuracy metrics (Radoux and Bogaert, 2014; Tyukavina et al., 2025). Alternatively, sample selection could be based on unit area to yield a sample with **inclusion probabilities proportional to unit areas**. These unequal probabilities should be accounted for at the analysis stage by using

estimators that explicitly incorporate the inclusion probabilities of the sample units (see Chapter 5 of the current protocol and Tyukavina et al. (2025)).

An **area frame** defines the geographic extent of the population boundaries from which the sample is selected. Area frames are useful when it is **impossible to list all sampling units in a population**, e.g., when sampling points from a continuous population. Continuous point sampling is the most common application of the area frame. An area frame is also useful when listing all sampling units is **impractical**. For example, suppose we want to sample polygons defined by the reference classification ('reference polygons'). A list frame approach would require delineating all reference polygons in the population which would amount to a census of the reference population (Stehman and Wickham, 2011). Because such a census is impractical, selecting points from an area frame is used, and the polygons intersected by a sample point are selected for the sample, resulting in a polygon sample with inclusion probabilities proportional to polygon areas.

3.3 Common probability sampling designs

The protocol by which sampling units are selected into the sample is referred to as the **sampling design** (Stehman and Czaplewski, 1998). Design-based assessments of map accuracy are based on **probability sampling designs**, defined in terms of **inclusion probability**, which is the probability of a particular sampling unit to be included in the sample. The defining feature of probability sampling designs is that **inclusion probabilities for all units in a population are greater than zero, and** inclusion probabilities **are known** for the units selected in the sample (Stehman and Czaplewski, 1998).

Examples of common probability sampling designs are **simple random, stratified random, systematic** and **cluster sampling** (see Figure 3.3.1). Systematic and cluster sampling can also be stratified, and cluster sampling can be implemented using a simple random or systematic selection protocol. Stratified systematic design, with varying per-stratum grid size densities, is more challenging to implement than stratified random (Stehman, 2009a), and hence, is less common. **When implementing these standard sampling designs in practice, the inclusion probabilities do not have to be computed separately, because they are already incorporated in the standard estimation equations** (Stehman and Czaplewski, 1998).

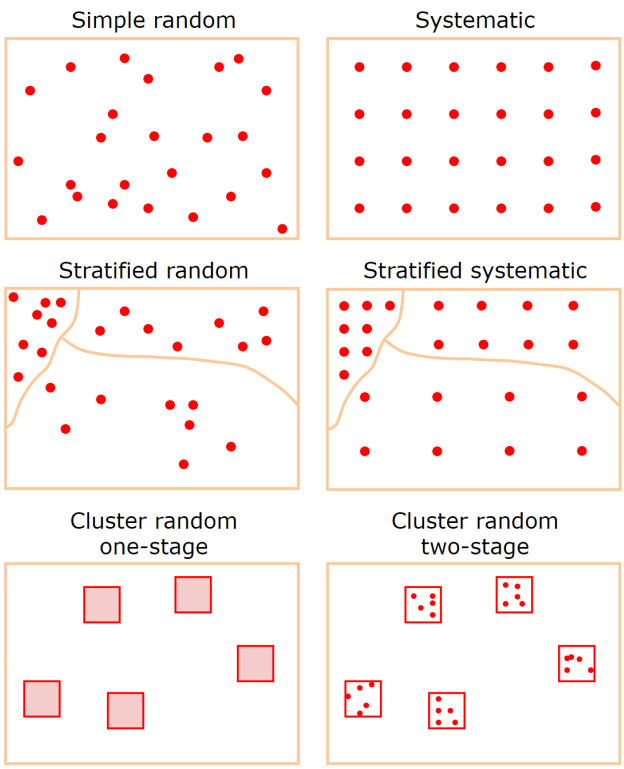

Figure 3.3.1 Common probability sampling designs. Red dots represent individual points or pixels and red boxes represent blocks of pixels. Stratified random and systematic designs are shown with varying sampling density across strata; cluster sampling with simple random sampling both in the first and the second (in case of two-stage) stages. Filled squares, in one-stage cluster sampling, indicate that all units within the cluster have reference labels assigned. Figure by Alexandra Tyukavina and Anna Komarova.

Non-probability sampling includes selecting reference data from conveniently accessible sites (e.g., along the roads) or using only locations where (non-random, non-systematic) aerial photography or high-resolution imagery are available. Such data collection protocols result in some areas in the target region having an inclusion probability of zero; e.g., areas further from the road or not represented in the available high-resolution imagery. Thus, the data collected in such a way **does not represent the entire target region of the map, and design-based inference cannot be used** (allowing the alternative but less satisfactory option of defining the more limited population of inference consisting of units that have a non-zero inclusion probability). Instead, non-probability sample data should be used in a **model-based inference framework** (see examples of model-based local map quality assessment methods in section 7.4).

When **selecting a probability sampling design**, three basic design choices need to be made (Stehman, 2009a):

1) whether to use a **simple random or systematic** selection protocol;

2) whether to use **strata**;
3) whether to use **clusters**.

The first question addresses the randomization protocol for selecting units into the sample, while the latter two address if and how to group the units prior to selecting the sample (Stehman, 2009a). Answers to these questions **depend on the study objectives and various practical concerns**.

Both **simple random** and **systematic sampling** are **equal probability sampling designs**, meaning that each element of the population has an equal probability of being selected in the sample. **Both sampling designs permit unbiased estimators of accuracy and area.** Simple random sampling is readily applied to select a sample of clusters, a sample of units within a cluster (two-stage cluster sampling), or a sample of units within a stratum (Stehman, 2009a). Systematic sampling achieves spatial balance and therefore tends to produce better precision than simple random sampling. Systematic sampling can potentially lead to poor precision of the estimates if the classification error is spatially periodic and the sampling interval coincides with error periodicity, but this combination of misfortunes should occur very rarely. Another disadvantage of systematic sampling is that it is not possible to construct an unbiased estimator of variance (standard errors) and therefore an approximate variance estimator must be used in practice (Stehman, 2009a). Systematic sampling allows the option of increasing the sample size by a specified number of units, but with some caveats. A systematic sample can be densified by decreasing a systematic grid distance to half the original distance, but this leads to a 4-fold increase in sample size which might be much larger than the sample size desired (Stehman, 2009a). Decreasing the density of the systematic sample (e.g. by doubling the grid distance) has the similar result of decreasing the sample size four-fold. An obvious option allowing flexibility of the increase or decrease in sample size would be to add sample units completely at random or to remove sample units completely at random, but this option disrupts the intended equal spacing of systematic sampling and consequently diminishes some of the advantages that accrue to systematic sampling. Sampling designs based on simple random sampling (stratified and non-stratified) allow adding or removing the exact number of sampling units to achieve desired sample size, if the process of selecting or removing the units is strictly random.

The second question is **whether to use strata**, which are *"subpopulations that are non-overlapping, and together comprise the whole population"* (Cochran (1977), p. 89). As **stratification requires assigning all units in the entire target region to a specific stratum**, stratification is not a viable option when point sampling is used to select reference polygons (see section 3.2), because constructing strata would require a census of reference polygons, which is not practical (Stehman and Wickham, 2011). Stratification is **useful when a small proportion of the population is of interest** (such as rare land cover classes or small geographic areas), as stratified sampling can ensure sufficient

sampling in those classes or areas. Stratification can also be used to control costs, e.g., when stratifying by distance from a road. In stratified designs, **sampling within a stratum is performed independently of sampling within other strata**, which allows the option of tailoring the sampling design for specific objectives or practical constraints in each stratum (Stehman, 2009a). For more discussion related to stratification please refer to section 3.4.

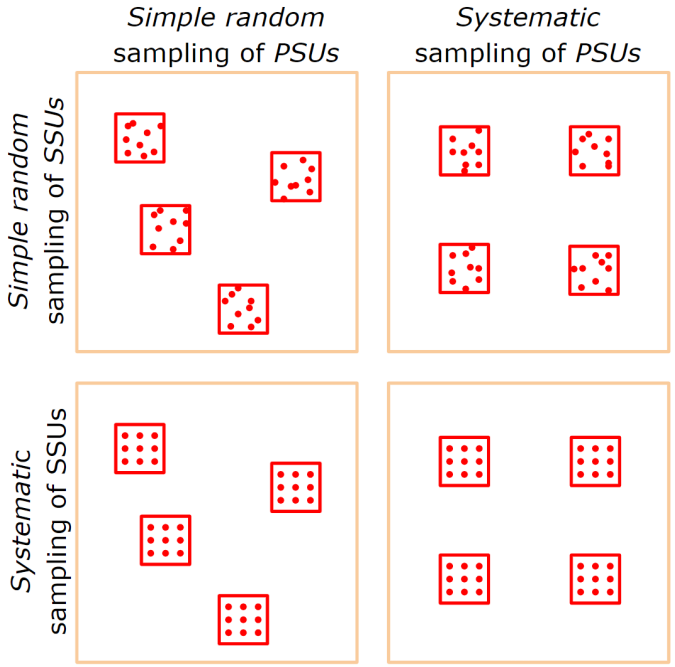

Figure 3.3.2 Common implementations of two-stage cluster sampling. Red dots represent individual points or pixels. PSUs stand for 'primary sampling units'; sampling of PSUs corresponds to the first sampling stage. SSUs stands for 'secondary sampling units'; sampling of SSUs corresponds to the second sampling stage. Figure by Anna Komarova and Alexandra Tyukavina.

The third question is **whether to use clusters**, which are groups of sampling units (e.g., 3x3 or 10x10 pixels). The blocks (clusters) are called primary sampling units (PSUs), and the units that form the clusters are called secondary sampling units (SSUs) (Stehman, 2009a). In **one-stage cluster sampling**, all units within each sampled PSU are included in the sample (e.g., all pixels within a pixel block are labeled). In **two-stage cluster sampling**, a second-stage sample of SSUs is selected from the PSUs selected in a first-stage sample. The most basic implementation of two-stage cluster sampling is selecting both SSUs and PSUs via simple random or systematic sampling (Stehman (2009a), Figure 3.3.2). The main benefit of cluster sampling is reduced cost, as reference data are required only for selected PSUs and not the entire study area. The potential disadvantage of cluster sampling is that it may yield larger standard errors compared to an unclustered design of the same cost, depending on the correlation among SSUs within

the PSUs; i.e., within cluster correlation (Stehman, 2009a). Two-stage cluster sampling typically has reduced reference data collection costs relative to one-stage sampling while reducing the effect of variance inflation due to the positive within-cluster correlation of classification error (Stehman, 2009a). The analysis of sample data collected under cluster sampling tends to be more complicated, particularly for two-stage cluster sampling. Also, it is more challenging to implement stratification based on the map classes of the SSUs (e.g., pixels) for cluster sampling (Stehman et al., 2008).

Stehman (2009a) provides a comparison of common probability sampling designs based on the following criteria: practicality (ease of implementation), cost effectiveness, spatial balance, sampling variability (standard error), availability of estimators that do not rely on approximations other than those related to sample size, and the ability to accommodate a change in sample size at any step in the implementation of the design. **Stratified random sampling** has the highest combined score based on these criteria and **is a recommended 'universally adequate' general purpose sampling design** (Olofsson et al., 2014; Stehman, 2009a).

3.4 Stratification

As mentioned in the previous section, **stratification partitions the area of interest into mutually exclusive subsets** (Olofsson et al., 2014), so that **each unit of the population belongs to only one stratum**. Stratification **can be based on the map being validated or an auxiliary variable** that is correlated with the target land cover class, which is useful for increasing the precision of the accuracy and area estimates by reducing variance within each stratum (making strata more homogeneous). For example, a smaller but thematically important land cover change class can be separated into its own stratum; the resulting 'stable' and 'change' strata will likely each have smaller variance of the target class values compared to simple random sampling within the entire target region. For multi-year land cover maps, individual mapped years or epochs can be used as map strata, especially if the goal is to estimate map accuracy for each year of epoch. **Sampling in each stratum is performed independently**, so a **small but important stratum can have a higher sampling density** (more units sampled per area) compared to larger stable land cover strata. A common goal of stratified sampling is to increase the sample size in small strata which otherwise might have few sample units selected if the design is simple random or systematic.

For estimating the area of relatively rare land cover or change classes, **additional buffer strata can be created** around the mapped land cover class (Figure 3.4) **to mitigate the effect of target class omission errors** on the precision of the estimates (Olofsson et al., 2020). Omission errors are typically located close to the boundary of mapped classes (i.e., mixed pixels), and therefore using buffers is often effective to

spatially isolate the areas with higher omission error rates into a separate stratum. Olofsson et al. (2020) discuss buffer size selection for estimating the area of deforestation.

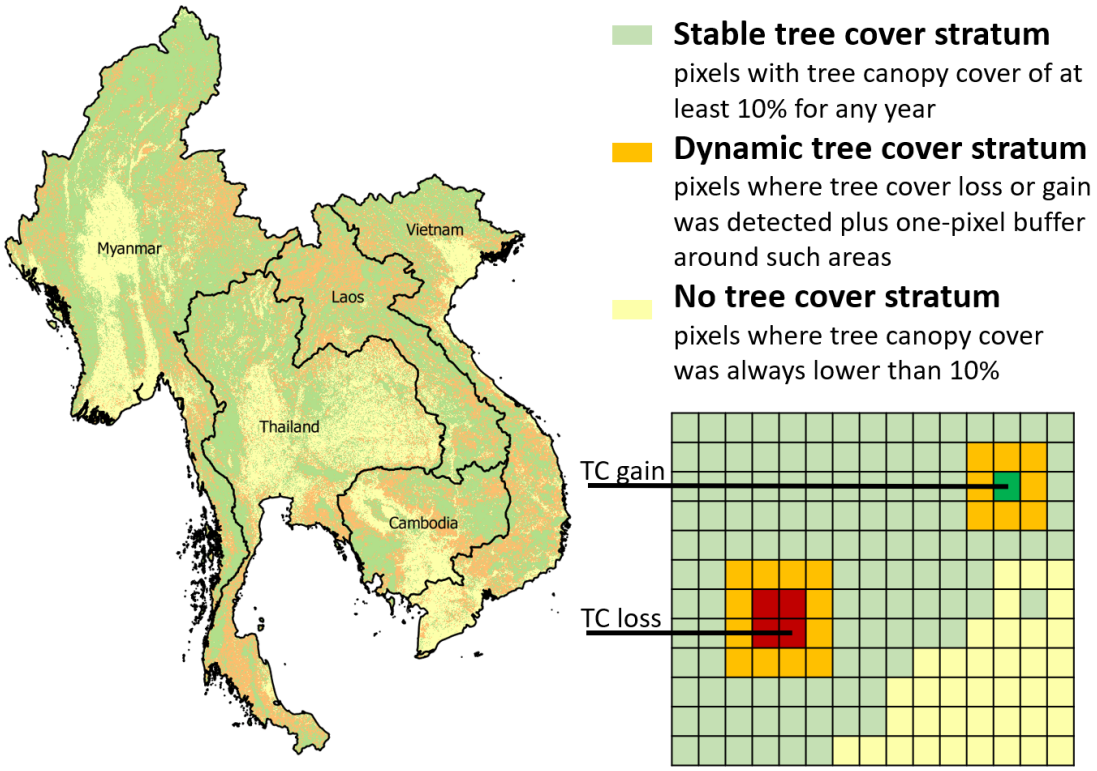

Figure 3.4 Example of stratification targeting potential omission errors of a rare land cover class (tree cover change) using a buffer stratum (one pixel buffer around mapped tree cover change). TC stands for tree cover. Figure by Peter Potapov.

At the same time, **stratification does not have to be based on the map being validated**. For example, if a goal is to report map accuracy by country, national boundaries could be used as strata (in this case the strata are often referred to as 'reporting regions' to distinguish from map-based strata, although the two types of strata are the same in terms of the analysis). Other examples of reporting region strata are continents, biomes, protected areas or emission factor-based strata within which activity data areas for carbon reporting need to be estimated (IPCC, 2006).

The **map-based strata do not have to match the map being validated**. Thus, the stratification can be defined based on one map or a combination of multiple maps and later used to assess the accuracy of multiple maps. While doing so, it is **important to keep track of the original strata**, their weights (stratum size relative to the entire target region) and sample sizes, since the **strata become a fixed feature of the design once the sample is selected** (Olofsson et al., 2014). For data collected under simple random or systematic sampling, post-stratified estimators can be implemented by stratifying the

study area after sampling (**post-stratification**). Post-stratification, which is a model-assisted estimation option (see section 5.2), can offer benefits of increasing precision of the estimates by partitioning the original simple random or stratified sample into more homogeneous post-strata (with smaller within-stratum variance), even without increasing the sample size within these post-strata.

While it is often beneficial to create a separate stratum for a small land cover or change class, issues could arise with multi-class land cover and change maps as the number of strata will be large. **The larger number of strata requires an increasingly larger sample size** (assuming we want to maintain a minimum sample size per stratum) and, consequently, **increased sample interpretation costs**. Therefore, in practice it is often beneficial to **combine multiple smaller land cover or change classes into more generalized strata** (Olofsson et al., 2014; Stehman, 2009a)**.** This is especially relevant when creating a stratification based on a multi-class land cover change map, in which the number of transition (from-to) classes could be very large, but the area of each class could be relatively small and the map accuracy for small transition classes relatively low, thus reducing the effectiveness of stratification based on each individual transition class.

Finally, estimates are often required for subsets of the original sampling population (**reporting domains** or **subdomains**) such as individual countries when sampling globally, or regions within a country when sampling nationally. If the domain corresponds to a stratum or a combination of several strata, then the usual stratified estimators can be applied. A broader discussion of domain estimation is beyond the scope of this document and we refer the reader to Chapter 10 of Särndal et al. (1992) for further details.

3.5 Sample size planning and allocation to strata

Addressing a common misconception, it is important to note that the **standard error of the sample-based estimates** of both area and accuracy metrics **depends on the absolute size of the sample**, and **not** on the percent of the population sampled. Cochran (1977), p.24 discusses this issue in relation to the finite population correction in standard error equations: "*Provided that n/N [where N is the size of the population, n is the size of the sample] remains low, these factors are close to unity [factors are (N-n) and N in the finite population correction, which is computed as (N-n)/N; when N is sufficiently large relative to n, this correction is close to one],* ***and the size of the population as such has no direct effect on the standard error of the sample mean***." From Lohr (2010), p.46: "*Instead of asking about required precision, many people ask, 'What percentage of the population should I include in my sample?' This is usually the wrong question to be asking.* ***Except in very small populations, precision is obtained through the absolute size of the sample, not the proportion of the population covered.***"

Two main questions of sampling design planning are, what **overall sample size** is needed, and **how to allocate the sample among the strata** when implementing a stratified sampling design. In the following, we focus on simple random and stratified random sampling, but the same planning strategy would apply to systematic sampling as well, albeit with the assumption that systematic sampling would yield similar precision to simple random sampling. To maintain the uniform grid spacing of systematic sampling, we note that systematic sampling is less flexible to adding or subtracting sampling units, and stratified systematic sampling with varying per-stratum grid densities is more challenging to implement in practice compared to stratified random sampling (see section 3.3 for more details).

Overall sample size planning

When planning the **overall sample size,** two scenarios are commonly encountered. In many applications the **project budget dictates an upper bound on sample size**, in which case the issue is whether that sample size will yield precision acceptable to project objectives. In other applications, formal sample size planning calculations provide the basis for **proposing a total sample size** and thereby specifying a major cost of the reference data collection. In either scenario, it is useful to go through the exercise of applying sample size planning formulas, and also to evaluate the anticipated standard errors that will be produced by a chosen overall sample size and sample allocation. Sample size and sample allocation planning provides the added benefit that the exercise requires deciding what the priority estimation objectives are because this information strongly influences sample design decisions. Although it is useful to conduct sample size planning, the more critical concern is the resulting standard errors (or widths of confidence intervals) achieved by the sample. These standard errors are, of course, not known until the project is complete. Sample size planning is always based on inputs that are not known with certainty (e.g., stratum-specific variances), so while the planning formulas give exact numerical outcomes, these outcomes should be regarded as guidelines.

There is also a related question of the **absolute minimum sample size** acceptable for the statistical accuracy assessment. The minimally adequate sample size will depend on the objectives of the assessment and number of classes. Given that estimating class-specific accuracy is included in the primary objectives of practically all accuracy assessments and that stratified sampling is well-suited to the objective of estimating class-specific accuracy, the rule of thumb proposed by Hay (1979) of **50 sampling units minimum per stratum** is reasonable guidance. Using Equation 3.1 provided below separately for each stratum, the standard error of estimated user's accuracy can be calculated for different values of user's accuracy and a sample size of 50 (e.g., if user's accuracy is 60%, the standard error would be 7%, and if user's accuracy is 90%, the standard error would be 4%). Thus, if there are only two map classes and two

corresponding sampling strata, then a total sample size of 100 may be sufficient if the standard errors achieved by a sample size of 50 per stratum are acceptable. If the **sampling design is not stratified**, a simple random or systematic sample would likely require a **larger sample size** so that the **minimum sample size per land cover class is approximately 50**. For example, a class that occupies 5% of the study area would, on average, have only 5 sample units in a simple random or systematic sample of n = 100, and call for the larger minimum sample size of n = 1000 in non-stratified sampling designs. In this case, when the target land cover class is relatively rare and the required minimum sample size is large, it is recommended to switch to a stratified sampling design with one of the strata targeting that rare class.

An important caution when the sample size is small is to recognize that **absence of omission or commission errors in the sample is not necessarily strong confirmation of absence of such errors in the map**. When no commission errors occur in the sample, the conventional standard error of the estimated commission error (and user's accuracy) would be 0, and it is not possible to compute the usual confidence interval based on adding and subtracting the standard error to the estimated commission error. The same issue applies to omission error estimates and producer's accuracy. In such cases an alternative confidence interval method must be used as the normal distribution methods are not appropriate. For example, for a simple random sample and applying the confidence interval method of Klaschka and Reiczigel (2021), the upper bound of a 95% confidence interval for the commission error percent when no commission errors are observed in the sample would be 2.3% if the sample size is 100, 4.5% if the sample size is 50, and 10.9% if the sample size is 20. The corresponding user's accuracy confidence lower bounds would be 97.7%, 95.5%, and 89.1% for sample sizes of 100, 50, and 20. The same method could be applied for omission errors and producer's accuracy in this example.

For **simple random sampling** and the objective of **overall accuracy estimation**, Stehman and Foody (2019) provide equation (equation 1) to estimate required **overall sample size** based on the **anticipated overall accuracy** and **desired confidence interval width**:

$$n = \frac{z^2 p(1-p)}{d^2}$$ Equation 3.1

Where n is the required sample size;
p is the anticipated overall accuracy (proportion of total area correctly classified);
z = 1.645 for a 90% confidence interval or z = 1.96 for a 95% confidence interval;
d is the desired half-width of the confidence interval (i.e., the confidence interval is estimated overall accuracy $\pm d$).

For example, if we anticipate overall accuracy to be p = 0.90 (90%) and we would like a 95% confidence interval (z = 1.96) to have a half-width of d = 0.02 (2%), Equation 3.1 yields a required sample size of n = 864. This same formula can also be applied when the objective is estimating the proportion of area based on the reference data. For estimating area, *p* is the anticipated proportion of area of the class of interest.

An alternative version of the above formula can be used if the goal is to achieve a specified standard error for overall accuracy or the proportion of area. The sample size needed to achieve a **target standard error** (equation 4.2 from Cochran (1977), p. 75) is:

$$n = \frac{p(1-p)}{SE^2} \qquad \text{Equation 3.2}$$

Where p is the expected proportion (either overall accuracy or area);
SE is target standard error expressed as a proportion.

For example, for estimating proportion of area, if we expect the proportion of the target class p = 0.03 or 3% (rare class, e.g., forest loss); and our target SE is 0.003 (i.e., if the SE is targeted to be 10% of the target class proportion, then target SE is 0.03 x 0.1 = 0.003). Equation 3.2 then yields a required sample size of n = 3,233. In contrast, estimating the area of a common land cover class with expected class proportion p = 0.50 or 50% (e.g., forest area) with the same target SE of 10% of the target class area, would require only n = 100. These two examples reflect the fact that even though in both cases the target standard error is 10%, when expressed relative to the anticipated proportion p, achieving the much smaller absolute standard error of 0.003 requires a much larger sample size than achieving the standard error of 0.05.

Sample allocation among the strata and overall sample size planning for stratified sampling

For **stratified sampling**, the **overall required sample size** depends on the selected **sample allocation among the strata**, which is discussed below for different estimation objectives. Common types of sample allocation among the strata are **equal** (the same sample size in each stratum), **proportional** (allocation proportional to stratum area) and **optimal** (minimizing variance). Stehman (2012) explored the impact of sample size allocation when using stratified random sampling for different objectives and found that when working with two strata (e.g., 'change' and 'no change'), **Neyman's optimal allocation** (Equations 3.3-3.4) **was preferable when aiming to estimate overall accuracy or area of change**.

For the objective of **area estimation**, the optimal allocation equation uses conjectured or speculated values of the target class proportion in each stratum. The values of these per-stratum class proportions can be estimated from an initial small

sample allocated to each stratum (e.g., 50 units per stratum) if it is possible to perform multiple sampling iterations, for example, when sample reference labels are derived via visual interpretation of satellite imagery. The per stratum estimates from this initial sample will not be precise, but if no other data are available for the target class proportions needed for the sample size planning calculations, these initial estimates can be used. If it is not possible to do an initial pilot sampling, e.g., when sample reference labels are derived via a ground survey and it is impractical to perform multiple field trips, expected per-stratum class proportions can be based on prior knowledge about the nature of the strata and land cover classes being estimated.

The required **overall sample size** n, for estimating target class area in **stratified sampling for the optimal allocation** with fixed n, could be estimated using Equation 5.66 from Cochran (1977), p. 110:

$$n = \frac{\left(\sum_{h=1}^{H} W_h \sqrt{p_h(1-p_h)}\right)^2}{SE^2} \qquad \text{Equation 3.3}$$

Where H is the number of sampling strata;
$W_h = \frac{N_h}{N}$ is the weight of stratum h;
N_h is the size of stratum h (total number of units in that stratum);
N is the total population size (total number of units in the sampling region);
p_h is the proportion of target class in stratum h, estimated from the sample or guessed based on the prior knowledge;
SE is the target standard error of the class proportion, expressed as the proportion of the total area.

Optimal sample allocation of the total sample size n **among strata** could then be calculated using Equation 5.60 from Cochran (1977), p. 108 (minimized variance for fixed n approach):

$$n_h = n \frac{N_h \sqrt{p_h(1-p_h)}}{\sum_{h=1}^{H} N_h \sqrt{p_h(1-p_h)}} \qquad \text{Equation 3.4}$$

where n_h is the sample size that needs to be allocated in stratum h.

For the objective of **overall accuracy** estimation **in stratified sampling**, equations are the same as for area estimation (Equations 3.3 and 3.4), but the proportion of correctly classified units is used, instead of the proportion of the target class, for each stratum in the optimal allocation equations.

If the priority objective is estimating **user's accuracy** and the strata match the map classes, then allocating a sample size of 75-150 per stratum should be sufficient

(Equation 3.1 can be used for the sample size calculation). In this case Equation 3.1 is used to compute desired sample size n_h in each stratum separately (expected user's accuracy and its desired precision can vary by stratum), and the total sample size n is the sum of stratum-specific sample sizes n_h. Recall that the primary motivation for stratification to estimate user's accuracy is to increase the sample size for the rare classes. In practice, when working with multiple estimation objectives (e.g., to improve precision of producer's accuracy and area estimates in addition to user's accuracy), n_h values computed this way could be used as minimum per-stratum sample sizes; the sample size could be increased in larger strata to improve precision of other estimates as the standard errors of area and producer's accuracy are usually smaller when sample sizes are larger in common classes (Stehman and Wagner, 2024).

For **multiple estimation objectives**, Stehman (2012) suggests computing standard errors for **multiple allocation options between proportional and optimal** (for one of the estimation objectives) and choosing the allocation with most acceptable standard errors of all estimates. Wagner and Stehman (2015) provide a spreadsheet with a program to objectively determine the optimal allocation while simultaneously minimizing the sum of the variances of the estimates of user's accuracy, producer's accuracy, and area, but this approach is limited to a single target land cover class.

Overall sample size n for **proportional allocation** among the strata (sample units are allocated proportionally to stratum size (total number of units) or area) can be computed using equation 5.65 from Cochran (1977), p. 110:

$$n = \frac{\sum_{h=1}^{H} W_h p_h (1-p_h)}{SE^2} \qquad \text{Equation 3.5}$$

Please note that Equations 3.1-3.5 do not include a finite population correction, since the size of the population (sampling region) N is usually much larger than the computed sample size n in map accuracy and area estimation applications (population usually consists of millions of units, e.g., pixels), and thus n/N (or n_h/N_h in each stratum in the case of stratified sampling) is negligible.

In practice, minimizing standard errors of accuracy estimates (overall, user's and producer's accuracy) is rarely prioritized as an estimation goal, unlike the standard error of the estimated area of the target class. For example, international agreements, national or corporate commitments on reducing deforestation often have a **requirement for precision of the reported estimates of the area** of deforestation. Often this required precision is stated as a standard error of the estimate, expressed as a percentage, e.g., standard error not exceeding 10% of the area estimate. In this case (single estimation objective, clear standard error goal), and when using Neyman's optimal allocation, it is

relatively easy to compute the required overall sample size (see Equations 3.3-3.4 and numerical example below). For **multiple land cover classes**, the required overall sample size and optimal allocation among the strata could be computed for each class and its required precision level separately, and then decisions on prioritizing each class could be made based on the required sample sizes in each stratum, and on the feasibility and labor costs for deriving sample reference values. For example, optimal allocation for estimating the area of land cover class 1 suggests allocating 100, 200 and 500 units in strata 1-3, respectively, and optimal allocation for estimating the area of land cover class 2 suggests 300, 150 and 700 units in strata 1-3. Then a maximum of required sample sizes in each stratum could be taken to reach required precision for both classes (300, 200 and 700 units in strata 1-3), if the resulting overall sample size of 1200 units is feasible in terms of reference data collection costs. If it is not feasible, decisions aimed towards reducing overall sample size need to be made, such as de-prioritizing one of the land cover classes (using optimal allocation for only one land cover class that is more thematically significant, e.g., allocation for class 1 in the example above, which will result in a total sample size of 800 units), or taking a maximum feasible sample size (e.g., 600 units) and allocating it among the strata in the same proportion as the maximum of the two optimal allocations (150, 100 and 350 in strata 1-3). Although such decisions are subjective, they need to be reported in the sampling design description.

To summarize a common use case, **the algorithm for estimating the required sample size for stratified sampling** and estimating the **area** of a land cover class (or multiple classes), map **overall accuracy** or **multiple estimation objectives** includes the following steps:

Step 1a: Allocate a small initial number of sample units in each stratum (e.g., 50), interpret the sample, and based on the reference values for the target class (or classes) in each stratum, compute target class proportions or the proportion of correctly classified units (for overall accuracy).

Step 1b: If 1a is not possible, guess the target class proportions or proportions of correctly classified units from prior mapping and sampling experience.

Step 2: Compute the total required sample size (see Equation 3.3 above for optimal allocation and/or Equation 3.5 for proportional allocation if working with multiple estimation objectives) or identify the overall feasible sample size.

Step 3: Based on the total required or feasible sample size, compute optimal sample allocation among sampling strata, or evaluate multiple allocation scenarios (e.g., various options between proportional and optimal) to determine the best allocation option to reach a compromise between estimation objectives. The same estimated or guessed per-stratum target class proportions from steps 1a or 1b could be used as a basis for

calculating standard errors that would result from different allocations and these standard errors used to guide the decision of which allocation to implement.

Step 4: If the initial pilot sampling was performed, compare the existing sample allocation with desired allocation and add sample units to the strata that have a smaller sample size compared to desired. If no initial sampling was performed and sample size allocation decisions were based on the prior knowledge and guessed class proportions, allocate required sample size from step 3 in each stratum.

Numerical example

Step 1a: Initial pilot sample of 100 units has been allocated in each of the six sampling strata targeting forest loss in Nepal (600 units total), and reference labels of the target class (yes/no forest loss) were assigned to each sample unit (step 1a above, Table 3.5). Based on that initial sample, the preliminary estimated mean proportion of forest loss (from the total area of Nepal) is 0.029457 with SE of 0.006071 (20.6%). Target SE = 10% of the target class or 0.002945673 of the total area. The goal is to estimate **how many additional units need to be sampled in each of the strata to reach the target variance**.

Step 2: Using Equation 3.3 (optimal allocation, stratified random sampling) and information from Table 3.5, the estimated overall sample size is n = 1366.

Step 3: Using equation 3.4 the optimal sample sizes for the strata are: n_1 = 580, n_2 = 0, n_3 = 74, n_4 = 95, n_5 = 181, n_6 = 437. To avoid zero sample size in stratum 2 which would have compromised a probability sampling design (non-zero inclusion probability requirement for all population units) in the absence of the initial pilot sample of 100 units in that stratum, p_2 = 0 **should** be replaced with a small target class proportion. For example, with p_2 = 0.0001, the optimal sample size allocation among the strata is then n_1 = 571, n_2 = 22, n_3 = 73, n_4 = 93, n_5 = 178, n_6 = 430.

Step 4: Our initial pilot sample was sufficient in strata 2, 3 and 4, but more sample units need to be allocated to strata 1, 5 and 6 (471, 78 and 330 additional units using the allocation computed with p_2 = 0.0001).

Table 3.5. Data for the numerical example of calculating required sample size for estimating forest loss area (target class) in Nepal. For explanation of notation see Equations 3.3 and 3.4 above. Column 'Target class units' contains the number of sample units that were identified in the reference classification as a target class (forest loss). p_h is then estimated as the proportion of these target class units from the initial sample size n_h. Note that $p_h = 0$ in stratum 2 should be replaced with a small proportion (e.g., 0.005 or 0.0001) for the purpose of sample size planning, because the initial pilot sample of 100 units likely underestimates a very small proportion of the target class in that stratum, and a p_h of 0 will result in the optimal allocation of the sample size of 0 to that stratum, which would have compromised a probability sampling design (non-zero inclusion probability requirement for all population units) in the absence of the initial pilot sample of 100 units in that stratum.

	Stratum	***N_h***	***Initial n_h***	***Target class units***	***p_h***	***W_h***	***Final n_h***
1	stable non-forest	15583185	100	1	0.01	0.4648024	571
2	core stable forest	5997025	100	0	0	0.1788743	100
3	loss	552977	100	85	0.85	0.0164937	100
4	gain	843112	100	10	0.1	0.0251476	100
5	1-pix buffer around loss/gain	2213447	100	5	0.05	0.0660209	178
6	periphery stable forest (10-pix buffer inside)	8336728	100	2	0.02	0.248661	430
	Total	33526474					

4. Response design

In studies of land cover from remote sensing, an accuracy assessment is essentially an analysis to determine the magnitude of error in the class labels produced by an image classifier. The basis of the accuracy assessment is normally the comparison of the labels produced by the classifier against those observed in a reference classification for a selected sample of cases. The sampling design discussed above (Chapter 3) is focused on the selection of an appropriate sample of cases (e.g., pixels) to use for the accuracy assessment. The **response design** of an accuracy assessment **contains all the steps that lead to a decision regarding the agreement between the labels from the map and those in the reference classification for the selected sample** (Olofsson et al., 2014; Stehman and Czaplewski, 1998). Inappropriate construction or application of the response design always undermines the accuracy assessment results, while thoughtful response design may greatly improve reference data quality. It is therefore of **paramount importance to allocate the necessary resources and effort into the response design**.

The response design requires consideration of some fundamental issues. These include **1) selection of the appropriate assessment unit and spatial support unit** that links the land cover classification legend with the sampling unit (see definitions of sampling unit, assessment unit and spatial support unit below)**, 2) the definition of what constitutes agreement between the map and the reference sample classification, 3) the source and quality of the reference classification, 4) the independence between the map production process and its validation** and **5) the actions taken to address problems that arise in the response design and reference sample interpretation**.

A wide variety of categorical maps exist (Ahlqvist, 2005; Comber et al., 2005) that require specific response designs. The response design depends on the assessment unit and the classification system used to define the labels of the map (Radoux and Bogaert, 2017; Stehman and Foody, 2019). Here distinction needs to be made between **sampling unit**, **assessment unit** and **spatial support unit**. Sampling unit defines which units are selected in the sample, assessment unit defines the spatial scale at which the reference labels are assigned, and spatial support unit defines the area that is taken into account when assigning reference labels to the assessment unit (in the mapping context spatial support unit is often referred to as ‘minimum mapping unit’ or MMU, see section 2.3). For example, a map pixel could be used as all three units, if individual map pixels are sampled, assessed, and land cover class definitions do not contain criteria of minimum patch area that would require a spatial support unit larger than a map pixel. Larger landscape context (spatial support unit), e.g., 100 m^2 or 1 km^2, could be considered while labeling reference sample units, even if the sampling (and assessment) unit is a point or

pixel (Stehman and Czaplewski, 1998), e.g., to match the spatial support unit of the map (MMU). Or, a primary sampling unit in cluster sampling could be a block of pixels (e.g., 3x3 or 5x5 pixels), but the assessment unit for map accuracy assessment should still be an individual map pixel (all pixels within a block labeled in one-stage cluster sampling, or a sample of pixels labeled in two-stage cluster sampling). A block-level assessment could be performed (a block of pixels serving both as a sampling and assessment unit), e.g., for the purposes of estimating target land cover class area, but it should not be used for assessing the accuracy of a pixel map, as the spatial scale of such a block-level assessment is different from that of the map that is being delivered to the users (Czaplewski, 2003). Map polygons could be used both as a sampling and assessment unit if the legend is based on the polygon's spatial support (Radoux and Bogaert, 2017).

A fundamental assumption made in an accuracy assessment is that a gold standard reference data (i.e., the reference data are labeled perfectly, without error) is used. Rarely, if ever, is this assumption true, meaning that the **reference data used in an accuracy assessment contain errors** which will propagate through the assessment (see section 4.2 below for more discussion on the quality of reference data). Although assigning a reference label to each assessment unit may appear straightforward, the response design is fraught with challenges (Foody, 2013). Errors in the reference data can arise from sources ranging from human error (vigilance errors, systematic errors and estimation errors) (Radoux et al., 2020) through confusion due to episodic events (e.g., cloud shadow, fires, seasonal water), to deliberate errors (Foody, 2014; Halladin-Dabrowska et al., 2019). Furthermore, in highly dynamic landscapes, errors may be expected to increase as the time gap widens between the collection of satellite data used to create the map and collection of the reference data. Finally, geolocation errors can also impact the agreement between the map and reference data.

Because of the errors in the reference data, an accuracy assessment relying on an imperfect reference classification can result in substantial under- or over-estimation of accuracy metrics. The magnitude and direction of this bias is a function of the nature of the reference data and their association to the classification both in terms of quality and independence of the map production and validation protocols (Foody, 2024, 2023; Radoux and Bogaert, 2020). Methods to address the various concerns exist and it is sometimes possible to adjust an accuracy assessment to account for errors in the reference data (see section 4.2). However, these methods only partly reduce the impact of reference data errors. Thus, **response design should provide clear rules** determining the reference labeling protocol and establish reference data quality checks in the effort **to minimize** these **errors**. In a sense, response design should be a ‘how-to’ manual for the interpreters acquiring reference labels from the satellite imagery or collecting data in the field, including definitions of typical cases and rules on how to deal with uncertain situations.

A sound description of the response design, including **the evaluation protocol** (source and date of the information, definition of the assessment and spatial support unit) and **the labeling protocol** (specific information collected from each source, decision rules applied to assign class labels), should be provided along with the description of the sampling design (Chapter 3) and the analysis (Chapter 5) of the validation results. In addition, the transparency principle of validation good practice should be satisfied by providing information such as 1) the characteristics of the interpreters generating reference data (i.e., their experience and training) and 2) whether interpreters were unaware of the map labels of the assessment units they were interpreting (Stehman and Foody, 2019) and were independent of the map producers. This information is necessary to ensure the highest possible quality of the reference dataset and communicate its potential limitations. We recommend that response design metadata are reported based on the suggested categories presented in Table 4.1.

We recommend sharing the unit-level reference labels for the purposes of independent review and verification of the quality of the reference data. In some cases, when a reference dataset is created with the intention of independent validation of multiple land cover maps produced by different entities, it is possible to limit sharing the unit-level reference labels to keep the validation dataset independent from production (training) of the maps intended to be validated. In these cases, this justification for not sharing the reference labels should be prominently stated, and other descriptive metrics of the quality of the reference data (e.g., interpretation certainty for each assessment unit or results of an assessment of agreement between multiple interpreters assigning reference labels to a subsample) should be reported instead.

Table 4.1. Suggested response design metadata to be reported to ensure transparency, reliability, reproducibility, and map relevance of the response design, and also to document quality assurance procedures. Stars indicate the level of diligence required to satisfy the specific criteria (one star - minimum level of diligence, three stars - highest level). †10% of reference labels coming from expert consensus is not a strict/objective quality control guideline, but an example of what quality control might look like when the interpretation protocol is not 100% expert consensus-based (e.g., reference data are derived via crowdsourcing, or each assessment unit is interpreted by only one expert).

Category/Criterion		Description
Evaluation protocol	Required	• Define assessment and spatial support unit; • List sources or reference data (see Chapter 6), their resolution, quality flags, pre-processing steps and any other relevant information characterizing the quality of reference data;

Category/Criterion		Description
		● Specify the dates of the reference data sources and whether there is any mismatch between the dates of the reference data sources and the map; ● Provide educational/professional background and specify the training received by interpreters generating reference data, both for expert-based and crowdsourced assessments (see section 6.5).
Labeling protocol	Required	● Define rules for applying classification legend to label each assessment unit and the rules for defining agreement between the reference classification and the map; ● Define specific information collected for each assessment unit from each reference data source; ● (if applicable) define sub-sampling protocol (rules of estimating proportions of land cover classes/components within assessment units).
Independence	*	The map producer validates their own product based on the set-aside of the training sample. This practice is only possible if the training dataset is a probability sample and not recommended as errors in the reference data will be correlated with the errors in training data, and, consequently, in classification result (map), which will likely result in overestimation of map accuracy (see section 4.2).
	**	The reference sample is different from the training sample used to create the map AND the interpreters creating reference data are unaware of the map class for the assessment units they are interpreting, BUT the entity performing the validation is NOT distinct from the map producer (e.g., the reference data and the map are produced by different people working on the same team, or the map producer(s) are a part of the team validating the map, but are unaware of the map labels for specific locations).
	***	The reference sample is different from the training sample used to create the map AND the interpreters creating reference data are unaware of the map class for the assessment units they are interpreting AND the entity performing the validation is distinct from that of the map producers.

Category/Criterion		Description
Quality of reference data	*	Reference data are not quality controlled AND (there is no evidence that the reference classification is based on higher quality data (e.g., field OR higher spatial or temporal resolution data) OR higher quality labeling procedure than the map classification).
	**	Reference data are quality controlled (e.g., 10%† of the reference labels are verified via expert consensus) OR (there is evidence that the reference classification is based on higher quality data OR higher quality labeling procedure than the map classification).
	***	Reference data are quality controlled (e.g., all reference labels are screened based on assigned certainty or interpreter disagreement, and lower certainty labels are verified via expert consensus) AND (there is evidence that the reference classification is based on higher quality data AND higher quality labeling procedure than the map classification).
Agreement with classification legend	*	The scale and spatial support unit of the classification legend are not explicitly stated.
	**	The scale and spatial support unit of the classification legend are explicitly stated, BUT the agreement between the map and the reference classification is defined at a different scale (mismatch between map unit and assessment unit OR between spatial support units of the map and the reference classification).
	***	The scale and spatial support unit of the classification legend are explicitly stated AND (the assessment unit is the same as the spatial support and mapping unit, OR the map and the reference classification have the same assessment/mapping unit and spatial support unit)
Geolocation errors	Yes/No	● Geolocation errors are described separately from the thematic accuracy (i.e., tolerated when building the confusion matrix). ● Geolocation errors are not distinguished from thematic accuracy (i.e., included as errors in the confusion matrix).

Category/Criterion		Description
Reference labels reported for each assessment unit	Yes/No	• Specific information collected for each assessment unit from each reference data source is provided, including the confidence of label assignment. • Unit-level reference labels are not provided.

4.1 Sample labeling protocol

The same set of rules must apply to the reference sample labeling protocol and that of the map classification system. Consequently, an ill-defined map legend (classification system) introduces uncertainty to the reference data. The **meticulous definition of mutually exclusive classes is thus critical to the development of the labeling protocol**. For example, the Land Cover Classification System (LCCS) proposed by Di Gregorio (2005) and its successor the Land Cover Meta Language (LCML) (Di Gregorio and Leonardi, 2016) provide the basis for a consistently applied legend (Olofsson et al., 2012).

The rules defining the agreement between the map and the reference data are dictated by the legend and the conceptual model of the map (e.g., mapping discrete entities, spatial regions/classes or continuous fields of land cover variables; for more discussion on conceptual map models see Burrough (1996) and Bian (2007)). The response design effort increases with the heterogeneity of the assessment unit, and hence often with its size. **The agreement between the reference classification and the map should be interpreted at the scale determined by the map legend**, otherwise an accuracy assessment unit that is different from the map unit would lead to a mismatch of validation results with the spatial scale of the map (Stehman and Wickham, 2011). A common misapplication related to mismatching assessment units, discussed in section 3.1, is determining the agreement between the map and reference labels at the block level in cluster sampling, instead of per-pixel level within each cluster, when evaluating the accuracy of a pixel-based map. Block-level accuracy assessment is possible (e.g., if reference data are only available at the aggregated scale), but in this case the accuracy of the map aggregated to the block level is assessed, and not the accuracy of the original pixel-based map. Similarly, Radoux and Bogaert (2017) recommend the use of the map's polygons as assessment units when the legend of the map is object-based.

Categorical maps: single category per assessment and map unit

The **categorical geographic data model** implies that each **map unit is assigned to a single category** and hence represented as homogeneous. However, for pixels, blocks or polygons, i) the precision and accuracy of the boundaries of the spatial units

may lead to overlap with one or more land cover classes; ii) patches smaller than the map's spatial support unit (or minimum mapping unit (MMU)) can be included in larger spatial units; iii) some land cover/land use classes (e.g., tree savannahs, urban areas) are composed of a set of elementary landscape components or might not have abrupt boundaries (e.g., ecotone zone along the forest edge). This results in **mixed or heterogeneous mapping units**. As the proportion of mixed pixels in a map increases with the pixel size (Radoux et al., 2020), this issue has more impact on the labeling protocol with polygons or coarse spatial resolution pixels than with very high-resolution pixels. Nevertheless, the **labeling protocol should include appropriate guidelines for the case of mixed assessment units** (pixels or polygons) because they may occur at any scale. Furthermore, the response design should consider the mapping unit and the spatial support unit (or MMU) of the map in order to determine if a pixel is correctly classified. If the map is generalized to remove patches below an arbitrary size according to specific requirements, the same rule (spatial support unit) should apply to the response design by considering surrounding pixels to determine if there is an agreement on the central pixel. On the other hand, if the map has both a mapping unit and a spatial support unit (MMU) of one pixel, then the surrounding pixels should not be taken into account quantitatively (beyond general landscape context) except for when the geolocation errors are evaluated.

Estimating proportions of land cover classes/components within map units

The **LCCS-based** (using fixed proportion thresholds for the elementary land cover components in a hierarchical framework) and **majority-based** (selecting the modal land cover category within each unit) **legends require an estimate of the land cover proportion inside each map unit**. When the sampling and assessment unit match the map unit, this estimate can be obtained with a **subsample of points** (e.g., Bey et al., 2016), or partitioning an assessment unit into a **grid of squares** (e.g., Laso Bayas et al., 2017) or into **irregular polygons** (e.g., Achard et al., 2002; Lamarche et al., 2021) based on higher spatial resolution data (e.g., ground surveys or higher spatial resolution imagery). In case of majority-based legends, it is sometimes impossible to identify a single majority class for each assessment unit. The labeling protocol should therefore include an additional rule to consistently assign a class when this occurs. Using an odd number of partitions avoids undetermined majority cases when there are only two classes in the assessment unit, and using a larger number of partitions allows to determine the proportions more precisely in case of close class frequency values. Figure 4.1 Illustrates the different methods of estimating subunit land cover class proportions, described below, using a mixed 30x30 m pixel as an example.

Irregular polygons provide spatially detailed estimates of land cover proportions inside the assessment unit. In practice, they can be obtained via labeling of homogeneous

patches through automated segmentation or per-pixel automated classification or manually drawn by the photo-interpreter. These methods are hardly applicable in case of sparsely distributed land cover elements that define a specific land cover type by combination (e.g., wooded savannah, open forest). In addition, the delineation of polygons can be subject to delineation errors, especially if automated (Radoux and Bogaert, 2017). As the impact of these errors on the final reference label is difficult to quantify, the precision required for the delineation usually makes this approach more time consuming than labeling points or square partitions within the unit. Over-segmentation reduces the risk of missing important boundaries in the landscape, but it increases the number of polygons inside each assessment unit, and hence the labeling effort also increases.

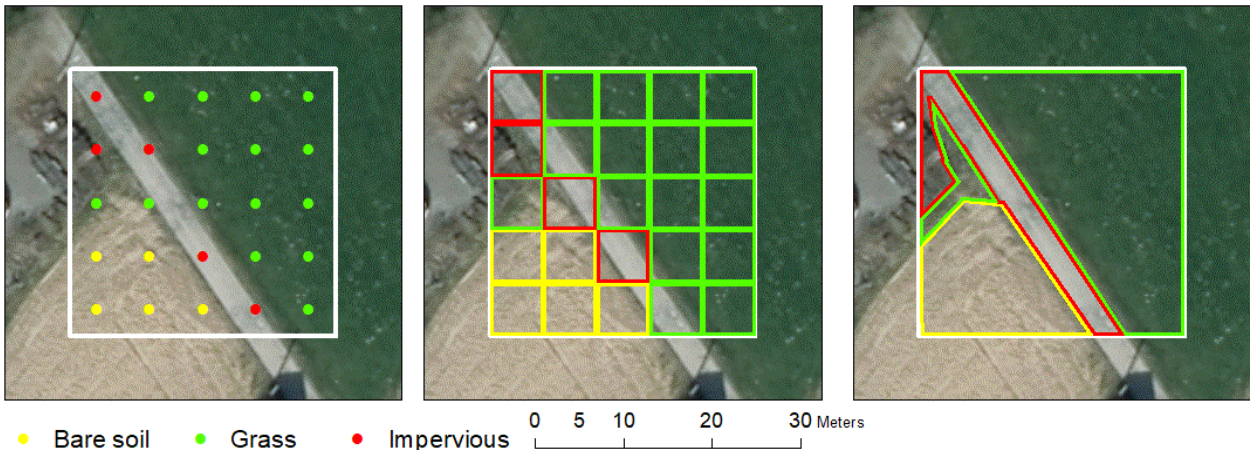

Figure 4.1 Estimating proportions of land cover classes inside a 30x30 m pixel **assessment unit** using a subsample of points (left) or partitioning an assessment unit into a grid of squares (center) or into irregular polygons (right). In this case, the three response designs would consistently assign 'grass' as the majority class with proportions of 60% with points, 64% with squares and 60% with irregular polygons. Figure by Julien Radoux.

Point-based (collecting information from a set of point locations within an assessment unit) and **square-based** (employing square partitions within an assessment unit) models for estimating subunit class proportions are easier to implement in validation workflows than polygons delineated by humans or machine learning. Their error rate depends on the number of subunits and on the frequency of mixed pixels with proportions similar to the class thresholds in the legend. For example, if the class 'forest' is defined based on a tree cover above 70% and there are a lot of mixed pixels with a proportion close to 70%, it is more challenging to build a high quality reference dataset. Block-based models are more accurate than point-based models for the same number of subunits when the legend is based on the majority, or when the class thresholds are close to 50% (Radoux et al., 2020).

As a way forward, **point-based subsampling could be optimized on the fly to adjust the number of points according to the required confidence on the labels for each assessment unit**. The accuracy of the labeling is lower when the proportions of land cover components inside the pixel are close to the predetermined thresholds. For example, if a mixed pixel is composed of 51% of class A and 49% of class B, a very high number of subsamples is required to find out the majority class with high confidence. But in the case of a pure pixel, it is useless to take more than a dozen subsamples to determine the majority class, and a human interpreter would assign the majority class without a doubt. Allowing some flexibility to the subsampling protocol is thus crucial to avoid unrealistic labeling effort. Some theoretical subsampling protocols allowed for a reduction of the labeling effort by 50 to 75% while providing an objective confidence level of labels for each assessment unit (Radoux et al., 2020).

4.2 Quality of reference data

A fundamental assumption typically made in accuracy assessment and area estimation is that the reference data used are an accurate representation of reality at the time represented by the map (Carlotto, 2009; Foody, 2010, 2002). That is, it is typically assumed in an accuracy assessment and/or in the estimation of class area that the reference data are a gold standard, absolutely correct in their labeling, and hence sometimes referred to as 'ground truth'. In reality this will rarely, if ever, be the case. **The reference dataset** is, like the thematic map under evaluation, merely a classification that **will contain some error**. The error may arise from a variety of sources ranging from **uncertainties in labeling**, through **human error** to **spatial and temporal mismatches between the map being evaluated and the reference data** (Foody, 2002; Powell et al., 2004; Thompson et al., 2007). A key issue, however, is that reference data error, while known to exist, is typically ignored, yet it can result in substantial misestimation of accuracy and area. Furthermore, the standard approach to accuracy assessment effectively attributes all errors to be in the land cover map and hence can be a somewhat pessimistic assessment of map accuracy (i.e., resulting in lower estimated map accuracy compared to its true accuracy, Foody, 2008). An optimistic bias (i.e., higher estimated map accuracy compared to its true accuracy) can also be introduced in map accuracy assessment depending on the nature of the errors in the reference data, in particular when the source of reference data is not independent of the map (Foody, 2024, 2023; Radoux et al., 2020). **Awareness of the existence of error in the reference data and its possible impacts can help enhance accuracy assessment and area estimation.**

Good practice advice for accuracy assessment urges that the **reference classification should be of a higher quality than the map under evaluation** (Olofsson et al., 2014). This could be achieved by 1) **obtaining higher quality reference data source** (e.g., authoritative data from the field collected by experts or remotely sensed

data of higher spatial, temporal and/or spectral resolution than the data used to produce the map) or 2) **if using the same source data for both the map and the reference classification** (e.g., Landsat imagery), **the method of deriving reference labels should be more accurate** than the mapping method (e.g., visual interpretation of satellite imagery for a limited number of sample locations is usually considered more accurate compared to automated classification of all pixels in the map). Ideally, **the same reference data source** (see Chapter 6) should be available **for all sample units**, and **the same labeling protocol** should be applied to all units to ensure consistency of reference classification; any deviations need to be disclosed in response design metadata (see Table 4.1)

However, **reference data error is** sadly **abundant** and often at levels that can be a major concern. Studies based on visual interpretation of remotely sensed imagery by multiple people often used as a source of reference data in mapping projects show that substantial error and uncertainty exists. For example, Powell et al. (2004) highlight that expert image interpreters disagreed on ~30% of cases; Johnson and Ross (2008) show up to ~40% disagreement in labeling; Thompson et al. (2007) show differences in labeling for some 4% of forest cases labeled. Pengra et al. (2020) observed an overall agreement of 88% between interpreters and class-specific agreement ranging from 46% for Disturbed to 94% for Water classes, with the more prevalent classes (Tree Cover, Grass/Shrub and Cropland) generally having greater agreement than the rare classes (Developed, Barren and Wetland). There may be **greater uncertainty with reference data contributed by non-experts**. For example, crowdsourced volunteered geographical information (VGI) has become a popular source of reference data as it offers the potential to acquire large, well-distributed and timely data, but it may be acquired by amateurs with variable skill levels and thus may provide labels of variable quality (Fonte et al., 2015). Section 6.5 provides further discussion on crowdsources vs. expert-based reference data collection methods.

Even very **small errors in the reference data can be a cause of major misestimation of accuracy and area** (Foody, 2013, 2010; Radoux et al., 2020). For example, Pontius and Lippitt (2006) highlight that error in the datasets used to form a confusion matrix may make it infeasible to detect, let alone measure, land cover changes. Foody (2013, 2010) provides a guide to the magnitude of the impacts of reference data error on accuracy assessment and area estimation. In one scenario of a binary classification with very accurate reference data (image classification overall accuracy = 90%, reference data overall accuracy = 95%), the estimates of class area, user's and producer's accuracy were 23.0%, 67.3% and 76.1%, respectively, when the actual values calculated with a gold standard reference data (overall accuracy = 100%) were 20.0%, 69.2% and 90%, respectively (Foody, 2010). The magnitude of misestimation varies depending on a range of issues such as the degree of error in each dataset and the

abundance of the classes. The impacts can also be particularly severe for rare classes. For example, taking an extreme situation for illustrative purposes, using a classification with an overall accuracy of 70% and reference data with an overall accuracy of 80%, the area of a class that actually covered 0.5% of the map would be estimated to be 20.3%, over 40 times greater than the reality (Foody, 2013). It is vital, therefore, that **the quality of the reference data be considered**. Note that in some cases, the negative effects of reference data quality on accuracy and area estimation are potentially correctable (Carlotto, 2009; Foody, 2009; Radoux and Bogaert, 2020).

There are **two main sources of errors** linked with the reference data quality: **geometric errors** and **thematic errors**. Geometric errors arise from geolocation shifts between the map and reference data. Sources of thematic error include differences in the map pixel and ground data collection unit size, the time gap between image and reference data acquisition, and labeling errors. Labeling errors might arise from poor interpreter training (in both expert-based validation and crowdsourcing, see section 6.5), automated reference label annotation (not recommended in this protocol), or from combining data from multiple sources of potentially variable quality to derive reference labels. Geometric errors are usually independent of the classification errors, but thematic errors are sometimes correlated, especially in case of ill-defined response designs (e.g., using the same interpreter(s) for both reference data collection and map training, or interpreters influenced by the knowledge of the classification result).

Errors in the reference data affect both the accuracy of the map and the area estimates. Typically, the **accuracy of the map** will be **underestimated when the errors in the reference data are independent** and **overestimated when the errors in the reference data are correlated** with map errors (Foody, 2024, 2023, 2010; Radoux and Bogaert, 2020). Furthermore, estimation of land cover class area with imperfect reference data could lead to substantial **misestimation of area**, especially for rare land cover classes (Foody, 2013). When comparing accuracies of multiple maps using a single reference dataset, errors in the reference dataset correlated with errors in one of the evaluated maps might lead to incorrect comparison results, whereas reference data errors independent from all evaluated maps, or correlated with all maps in the same way (reference data and all evaluated maps tend to misclassify the same locations), won't affect such comparisons. For example, Radoux and Bogaert (2020) created a set of simulated maps with overall accuracies, estimated using a trusted reference dataset (with errors independent from all simulated maps), ranging from 80 to 93%. They then created a set of reference datasets that were not independent, each having 50% of the errors correlated with one of the simulated maps. Each non-independent reference dataset identified the map with the correlated errors as the most accurate one of the set, despite up to a 13% difference in the actual overall accuracy of the maps that were evaluated (estimated using a reference dataset independent from all maps). This example

demonstrates the magnitude of the differences in map accuracies that might be obscured if comparative map accuracy assessment relies on a reference dataset with errors correlated with errors in one or more (but not all) of the evaluated maps.

The challenges associated with imperfect reference data and the ways to address them **are the subject of research**. The latter includes the potential of model-based approaches, such as latent class analysis, which has been used in other disciplines for activities similar to accuracy assessment and area estimation when a gold standard reference dataset is not available (Foody, 2012, 2010) and how to exploit the opportunities offered with Volunteered Geographic Information (VGI), including ways to usefully integrate it with authoritative data in rigorous estimation (Stehman et al., 2018).

4.3 Accounting for reference data errors

With a correctly implemented sampling design and unbiased estimators as described in Chapters 3 and 5, respectively, the standard deviation of the estimates can be reduced by increasing the number of reference sample sites. However, **increasing the size of the reference sample does not reduce the bias of the estimates originating from low quality response design and resulting reference data errors**. The estimates would in this case converge to the wrong target. When the bias associated with the reference data becomes larger than the confidence interval defined based on sampling variance, **validation efforts should focus on improving the response design instead of adding more points of the same quality**.

Accounting for thematic errors in reference data

When errors in the reference data are known, map accuracy indices can be adjusted *a posteriori* to improve the estimates. Existing methods allow to correct sensitivity (producer's accuracy of the target class or recall) and specificity (producer's accuracy of a non-target class) estimates under the assumption of conditional independence (i.e., the errors in the map classification and reference data are independent and there is no tendency for the classification and reference to err on the same cases) (Foody, 2010; Staquet et al., 1981). Radoux and Bogaert (2020) compared different methods to reduce the uncertainty of the confusion matrix. For thematic errors, this requires a gold standard reference dataset as a subset of the main reference dataset to determine if there is a correlation between the classification errors and the main reference dataset. A confusion matrix can then be reconstructed to reduce the impact of the errors in the larger but lower quality reference dataset (e.g., a large VGI dataset) based on a small (e.g., a few hundred points) but 'near gold standard' dataset. When considering the RMSE of the overall accuracy (OA) and not only the variance of the estimator, a small gold standard reference dataset (e.g., 100 sample sites with overall accuracy = 99%) outperforms a large basic reference dataset (e.g., 10,000 sample sites

with overall accuracy = 95%). **It is thus worth spending 100 times more effort on the response design than collecting 100 times more sample units**.

McRoberts et al. (2018) found that when a **simple majority interpretation** of multiple experts evaluating the entire sample is used as a final reference sample label to estimate land cover proportions, the bias increased with greater inequality in strata sizes, smaller map and interpreter accuracies, fewer interpreters and greater correlations among interpreters. Using a greater number of interpreters is one way to mitigate the effects of interpreter error on estimates, and a hybrid variance estimator presented in McRoberts et al. (2018) allows to account for the effects on standard errors when **the entire sample is interpreted multiple times** (once by each individual interpreter). Stehman et al. (2022) present a less labor-intensive method, incorporating interpreter variability into the land cover class proportion variance estimates when only a **random subsample of a reference sample is interpreted by multiple experts**. Xing and Stehman (2024) provide an even more cost-effective solution of incorporating interpreter variability into land cover area variance estimates, by utilizing **interpenetrating subsampling**, in which the full sample is partitioned into several nonoverlapping subsamples each evaluated by only one interpreter, thus **not requiring repeat interpretations**. The methodology of incorporating interpreter variability into the estimates presented in these studies applies to the map accuracy estimation as well.

An alternative to using a majority interpretation from a large number of interpreters as a final sample reference label and incorporating the interpreter variability into the map accuracy and area estimates, McRoberts et al. (2018) suggest using the **consensus interpretation approach**. In the consensus approach, interpreters first independently assign reference labels to the assessment units and then discuss units with initial interpretation disagreements until consensus regarding the final reference label is reached. This consensus approach has been widely adopted in land cover map validation (Bassine et al., 2020; Hansen et al., 2022, 2013; Potapov et al., 2022a; Tyukavina et al., 2022). The consensus approach allows minimizing the interpreter variability in the reference sample labels instead of incorporating it into the final variance estimates. The consensus approach could be supplemented with assigning primary and alternate reference labels (a second choice, in case of doubt) or assigning confidence flags (high/low) to each interpreted assessment unit. This allows to test the effect of the low confidence assessment units on the estimates and to organize a consensus interpretation round (to focus interpretation effort or additional reference data collection on lower confidence assessment units). Low confidence assessment units, however, should not be excluded from the analysis, and secondary labels in uncertain cases should not be purposefully used to decrease the number of cases with disagreement between the reference classification and the map, but rather should be used as a measure of uncertainty of reference classification.

Accounting for geolocation errors in reference data

Geolocation errors are discrepancies in location of features in the map and in the reference data, with their actual location on the surface of the Earth. **Map geolocation errors** can originate from the input data (e.g., imperfect georeferencing or orthorectification of satellite imagery used to produce a map) and from the classifier (e.g., semantic segmentation errors). **Reference data geolocation errors** can also originate from imperfectly georeferenced or orthorectified imagery, or from the uncertainty of the global positioning system (GPS) when collecting ground reference data. When the geolocation error in the map has a much larger amplitude than the geolocation error from the reference data (e.g., 30 m pixels validated with reference locations derived via differential GPS), it is common not to distinguish geolocation error in the reference data from map thematic errors. On the other hand, the impact of geolocation errors is usually greater for high-resolution mapping (< 30 m), and it is then useful to distinguish geolocation errors in reference data from map thematic errors. A constant geolocation shift in reference data has a variable impact on the accuracy of the different classes. Fragmented classes and items with a vertical structure are typically the most affected (Radoux and Defourny, 2007). In any case, the response design description should mention whether the geolocation errors (both absolute geolocation errors in map and reference data, discussed above, and geolocation errors in the map relative to the reference data) are included in the confusion matrix or described separately from thematic accuracy (Table 4.1). It is also worth noting that if the source of reference data is the same as the source of classification data (map geolocation errors relative to reference data are absent), and absolute geolocation errors in that source data are not assessed, geolocation errors in the resulting map will be underestimated/not evaluated by design.

The main option for accommodating reference class ambiguity arising from **map geolocation errors relative to reference data** is to **assign a primary and an alternate (secondary) reference label** (Sarmento et al., 2009; Stehman and Foody, 2019) (Figure 4.3.1). This highlights the value of collecting reference data with information on surrounding areas. However, the presence of the secondary label within the search distance of geolocation tolerance is not sufficient to determine that the map is correct. It is a first step to identify the assessment units that need further attention. The neighborhood must then be interpreted to distinguish thematic and geolocation errors (see examples in Figure 4.3.2). This method requires revisiting all the assessment units for which the secondary label matches the map while also displaying the map labels, which is otherwise not recommended.

Alternatively to assigning primary and secondary labels, it is possible to build a co-occurrence matrix based on the probability of the assessment unit to fall on another class in a given neighborhood in order to correct the confusion matrix based on a sample with geolocation errors. On a synthetic use case with a map of 93% overall accuracy (OA) and

a geolocation uncertainty of one pixel, the observed OA was underestimated by 7.6% because of geolocation errors. This bias could be reduced to 0.26% after an appropriate correction (Radoux and Bogaert, 2020).

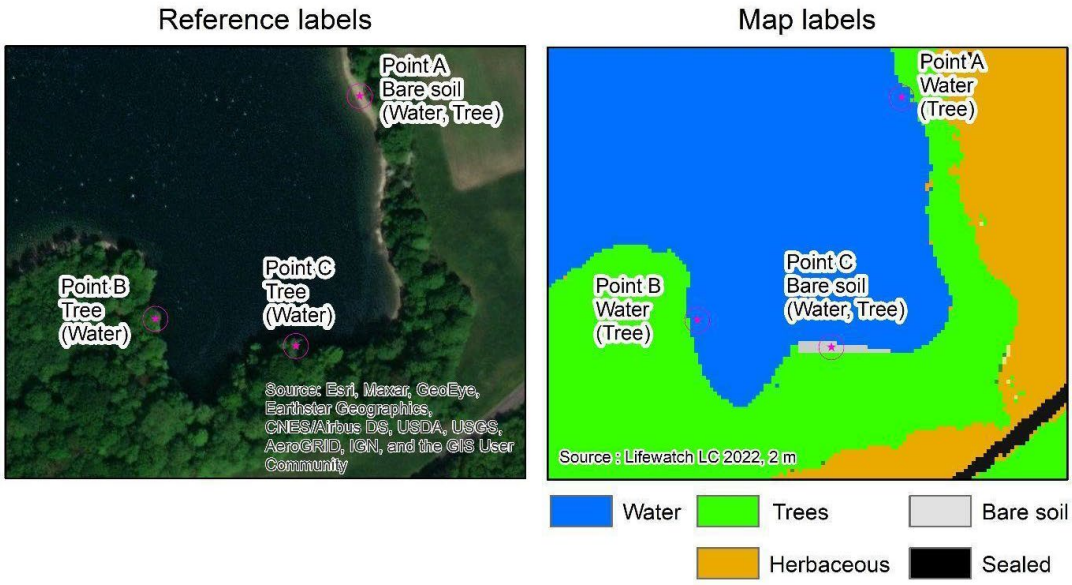

Figure 4.3.1 Examples of secondary labels (in parentheses) **used to handle geolocation errors** with the assessment unit of one pixel. The circle indicates geolocation tolerance of the assessment. A primary label (both in the reference data and in the map) is assigned based on the class that each point directly falls on, and secondary labels are assigned based on other classes within the circle. Point A has a primary reference label (bare soil) that is absent in the map within the circle. It should be thus considered a thematic error without further verification. Point B is along a boundary that is observable in the reference data and in the map. This can be considered a thematic agreement with a geolocation error. Point C is also along a boundary, but the primary map class (bare soil) is absent in the reference data. It should be considered a thematic error. Figure by Julien Radoux.

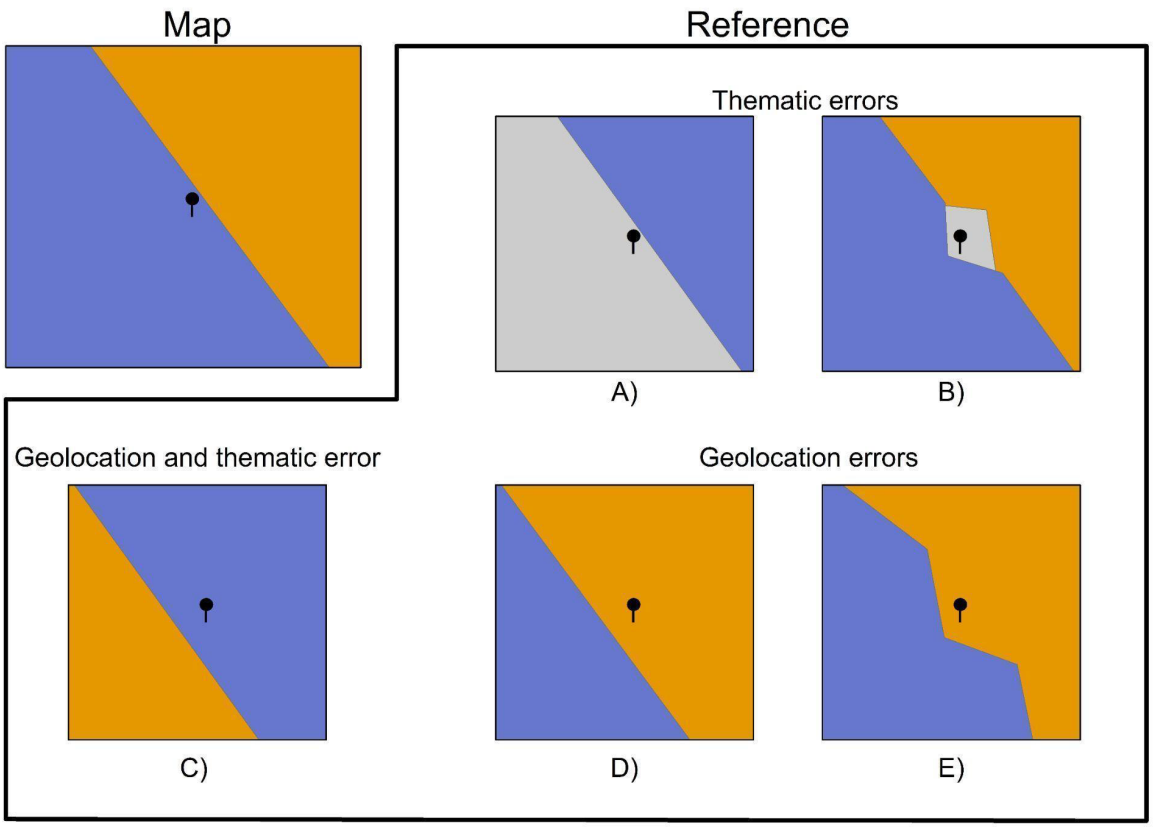

Figure 4.3.2. Example of situations where the secondary **reference label matches the primary label of the map, but the map is thematically incorrect** (A, B) vs. **when the secondary label might help identify geolocation errors in the absence of thematic error** (D, E). The black pin illustrates the assessment unit location, and secondary labels are assigned based on the entire square. A) and B) illustrate situations where the secondary label (purple class) matches the map's label, but the classification is actually incorrect. B) is a special case that could be accepted if the size of the gray class identified in the reference data is smaller than the map's spatial support (minimum mapping unit). D) and E) are typical geolocation errors. D is a geometric accuracy issue (systematic shift) while E is a precision issue (deviation around the correct position, which could be due to the finer spatial resolution of the reference data). Finally, C is a case that must be taken into account if location and thematic errors are processed separately. In this case, it should be flagged as both a thematic and a location error even though the two primary labels are matching. Figure by Julien Radoux.

5. Analysis

The analysis and estimation protocol is the third main component of a sample-based map accuracy assessment and area estimation, following a sampling and response design (Stehman and Czaplewski, 1998). The **analysis protocol** specifies the procedures to estimate characteristics of the study area from the reference sample data (Olofsson et al., 2014). **Design-based inference is the recommended analysis framework** (Olofsson et al., 2014) underlying estimation of population parameters from the sample data and the quantification of the uncertainty of these sample-based estimates. In design-based inference the **estimators (equations)** for map accuracy and land cover class area **depend on the selected sampling design** and the variance of an estimator is based on the set of all possible samples that could be selected by that sampling design (i.e., the frequency or randomization distribution of an estimator). It is recommended that only **unbiased estimators** are used for each sampling design (see section 2.1.4 for definition). All estimators presented in the current section are unbiased for a corresponding sampling design unless stated otherwise.

A **confusion** or **error matrix** is a tabular representation of correspondence between the map and the reference data (Figure 2.7, Table 5.1.1), where diagonal elements correspond to correctly classified map units, and off-diagonal elements are omission and commission errors. Typically, the population error matrix is not available, as wall-to-wall reference data do not exist, so a sample-based, estimated confusion matrix is used instead. A confusion matrix is a useful tool when computing and illustrating map accuracy measures (overall, user's and producer's accuracy), and is therefore suggested in many good practice guides (Olofsson et al., 2014, 2013; Stehman, 2013; Stehman and Foody, 2019).

When reporting a confusion matrix, **the elements of the confusion matrix should be expressed in terms of the percent or proportion of the total area**, and not in terms of the sample counts (Olofsson et al., 2014; Stehman and Foody, 2019). Under simple random sampling, the confusion matrix of the sample counts and of the area proportions both represent the area of the sampling region correctly. However, under stratified random sampling (see section 3.4), the sampling density is often not proportional to the size of strata, which necessitates an error matrix expressed in area proportions. Olofsson et al. (2014) and Stehman and Foody (2019) provide numerical examples for converting a confusion matrix of sample counts into area proportions.

Similar to the sampling design (Chapter 3) and response design (Chapter 4), the analysis protocol **should be described in a standardized and transparent way**, focusing on 1) the accuracy metrics associated with the land cover product and, if

applicable, 2) the area estimates of target land cover classes. We recommend that the analysis reporting is based on the suggested categories listed in Table 5.1.

Table 5.1. Suggested analysis components to be reported.

Category		Description of reporting elements
Accuracy metrics	Required	Confusion matrix with cell entries expressed as percent or proportion of the total area (Table 5.1.1); Overall accuracy; Class-specific accuracy metrics (user's and producer's accuracy for each land cover class at a minimum); Standard errors of all accuracy metrics; Employed estimators of accuracy (formulas used to compute accuracy metrics) are specified.
Area estimates	Yes/No	Optional, but recommended: sample-based area estimates for all land cover classes of interest, along with their standard errors (section 5.2); Employed estimators of area (formulas used to compute area estimates) are specified.
Local map quality metrics	Yes/No	Optional: information about local quality of the map (see section 7.4)

5.1 Estimating map accuracy

An **accuracy assessment** provides **estimates of the agreement between a map and a reference classification** (see definitions in section 2.1.1). The recommended approach of assessing map accuracy is to implement a probability sample with reference class labels obtained independently from the map and using a more accurate labeling method (Olofsson et al., 2014). Thus, map accuracy assessment identifies how well the map represents an independently derived reference dataset. See sections 2.2, 2.6 and 2.7 for more discussion on the general principles of estimating map accuracy.

The most widely used map accuracy metrics in remote sensing are overall, user's, and producer's accuracies. **Overall accuracy** is the proportion of the study area correctly classified (Stehman, 2014). In a confusion matrix, overall accuracy is derived from the diagonal elements of the matrix (Table 5.1.1). Overall accuracy is a single metric characterizing the entire set of map classes; it can be computed for the entire map or for

geographic subregions. Overall accuracy is perhaps the most commonly reported map accuracy metric, although it has limited utility for assessing the quality of the map as it does not provide class-specific accuracy information. Overall accuracy could be particularly uninformative when the class of interest is rare or when one class dominates the map. For example, if a land cover mapping project has two target classes, 'change' and 'no change', and the true area of the change class is small, let's say 5%, the overall accuracy of the resulting map is then 95% even if the whole area is classified as 'no change' (i.e. the change class is not mapped at all). The kappa coefficient was proposed in part to resolve this feature of overall accuracy, but **use of the kappa coefficient of agreement is strongly discouraged** by current good practice guidelines. Kappa is not informative because it is based on incorrect assumptions and is highly correlated with the overall accuracy, which makes it redundant (Olofsson et al., 2014; Pontius and Millones, 2011; Strahler et al., 2006).

As opposed to overall accuracy, user's and producer's accuracy express the accuracy of individual map classes and are therefore referred to as **class-specific accuracy metrics**. As such, these metrics can be more informative as indicators of map quality compared to the overall accuracy, as they provide insights into class-specific errors. At the same time, both overall accuracy and class-specific metrics quantify map accuracy within the entire mapping region or subregion, and thus they do not characterize local variations in map accuracy (see section 7.4).

User's accuracy (UA) represents the area of the class of interest that is correctly classified in the map, divided by the total area mapped as this class (Table 5.1.1). In other words, UA corresponds to the area of correctly mapped target class (true positive or TP), divided by the sum of TP and false positive (FP) detection map areas for this class (UA = TP / (TP + FP), Figure 2.7). The complement of UA is the commission error rate (1 - UA), representing the proportion of the mapped target class incorrectly classified as such.

Producer's accuracy (PA) is the proportion of area correctly mapped as the target class divided by the proportion of area identified as the target class by the reference classification (Table 5.1.1). In other words, PA corresponds to the area of correctly mapped target class (TP), divided by the sum of TP and false negative (FN) detection map areas for this class (PA = TP / (TP + FN), Figure 2.7). The complement of PA is the omission error rate (1 - PA), representing the proportion of the target class omitted, or not identified by the algorithm, in the map.

Overall accuracy of the map and user's and producer's accuracy for each map class **should always be reported with** their respective **standard errors** (Table 5.1), to demonstrate the precision of these accuracy metrics. Standard errors of accuracy metrics are an indicator of the sufficiency of the implemented sample size to precisely characterize map accuracy (see section 3.5). Confidence intervals (see definition in

section 2.1.4) of the estimated accuracy metrics at the required confidence level (e.g., 90% or 95%) can be computed from the reported standard errors.

Table 5.1.1 Error matrix of estimated area proportions, i.e. with cell entries estimated from the reference sample and expressed in terms of proportion of the total area, as suggested by good practice recommendations. The row ($\hat{p}_{i+}$) and column ($\hat{p}_{+j}$) totals are the sum of the $\hat{p}_{ij}$ values in each row and column. Adapted from Olofsson et al. (2014, 2013); Stehman and Foody (2019).

	Reference					
Map	**Class 1**	**Class 2**	**Class 3**	***Class 4***	**Total**	***User's accuracy of class i***
Class 1	$\hat{p}_{11}$	$\hat{p}_{12}$	$\hat{p}_{13}$	$\hat{p}_{14}$	$\hat{p}_{1+}$	$\hat{p}_{11}/\hat{p}_{1+}$
Class 2	$\hat{p}_{21}$	$\hat{p}_{22}$	$\hat{p}_{23}$	$\hat{p}_{24}$	$\hat{p}_{2+}$	$\hat{p}_{22}/\hat{p}_{2+}$
Class 3	$\hat{p}_{31}$	$\hat{p}_{32}$	$\hat{p}_{33}$	$\hat{p}_{34}$	$\hat{p}_{3+}$	$\hat{p}_{33}/\hat{p}_{3+}$
Class 4	$\hat{p}_{41}$	$\hat{p}_{42}$	$\hat{p}_{43}$	$\hat{p}_{44}$	$\hat{p}_{4+}$	$\hat{p}_{44}/\hat{p}_{4+}$
Total	$\hat{p}_{+1}$	$\hat{p}_{+2}$	$\hat{p}_{+3}$	$\hat{p}_{+4}$	1	
Producer's accuracy of class j	$\hat{p}_{11}/\hat{p}_{+1}$	$\hat{p}_{22}/\hat{p}_{+2}$	$\hat{p}_{33}/\hat{p}_{+3}$	$\hat{p}_{44}/\hat{p}_{+4}$		
Overall accuracy of the map	$\hat{p}_{11} + \hat{p}_{22} + \hat{p}_{33} + \hat{p}_{44}$					

User's accuracy and producer's accuracy are terms widely adopted in the land cover remote sensing community to refer to class-specific accuracy metrics. However, machine learning and data science literature uses other terms to describe the same metrics (Table 5.1.2). The user's and producer's accuracy terminology is so ingrained in the remote sensing lexicon that we recognize the likely continued use of these terms. However, it is necessary to be aware of these other terms and how to interpret them.

F1 score is a metric often used in the machine learning literature, which is the harmonic mean of the precision (UA of the target class) and recall (PA of the target class), computed as 2 x (UA x PA) / (UA + PA) (Christen et al., 2023; Emmert-Streib et al., 2019). F1 score can be computed for each class of a multi-class land cover map separately, making it a useful combined class-specific accuracy metric when performance of different maps for each land cover class needs to be compared using a single metric. For the objective of describing accuracy, the F1 score should always be accompanied by user's and producer's accuracy (or the complementary commission and omission error rates).

Table 5.1.2 Correspondence between terminology used for class-specific metrics of a binary map in Remote Sensing (RS), and Machine Learning (ML) and Data Science (DS) literature. ML and DS terms are from Emmert-Streib et al. (2019). UA stands for user's accuracy; PA stands for producer's accuracy. TP, TN, FP and FN represent cells of a binary confusion matrix corresponding to the area of true positive, true negative, false positive and false negative detections, respectively (see Figure 2.7 for a graphic representation of a binary confusion matrix).

	RS term	Equation	ML and DS terms
Focus on **target class**	UA	TP / (TP + FP)	precision, positive predictive value
	PA	TP / (TP + FN)	sensitivity, recall, true positive rate
Focus on **background class**	UA	TN / (TN + FN)	negative predictive value
	PA	TN / (TN + FP)	specificity, true negative rate

It might seem intuitive to obtain the areas of features on the land surface by counting pixels in maps, but because all maps have errors, pixel-counting is a biased estimator in the sense that it does not produce the true value due to these map classification errors (GFOI, 2016, p. 125). Instead, what is needed are areas with confidence intervals estimated from a sample of observations of reference conditions on the land surface (see section 5.2). However, it is natural to ask **what level of map accuracy is acceptable,** e.g., to justify use of pixel counting. There is **no universal answer** to this question. For applications where maps are used to implement land use interventions on the ground such as sending law enforcement to stop illegal deforestation, more conservative maps are more useful to avoid unnecessary effort expended at locations that were not really deforested. A conservative map underestimates the area of a land cover or change class (high omission errors, low producer's accuracy), but minimizes false detections (low commission errors, high user's accuracy). For applications that involve constructing map-based strata for sampling and that use satellite imagery interpretation as the primary source of reference labels, it may be useful to have a map with higher commission errors (lower user's accuracy), but low omission errors (high producer's accuracy). This is especially true if the phenomenon of interest is a very small proportion of the study area. In such situations, constructing a stratum with the aim of 'catching' or containing the omissions of said phenomena has proven efficient to reduce the variance of the sample-based area estimates (Olofsson et al., 2020). Having high and balanced user's and producer's accuracies (or low and balanced omission and commission errors; note that we are not referring to the number of errors but to the area proportion of error) results in a map that does not significantly over- or underestimate the target class area, which will be suitable for most applications. However, producing such a map might be impossible or impractical.

Recommended estimators of accuracy

Perhaps the best known and most widely used collection of estimators for map accuracy is presented in Olofsson et al. (2014). However, it represents a special case of **sampling from populations of equal-sized units** (e.g., pixel grids with constant pixel size) **with equal inclusion probabilities** of sampling units within each stratum, **when sampling strata match map classes**. Stehman (2014) provides a set of estimators that extends this special case to include situations when **sampling strata are different from map classes.** A further generalization of the estimation framework is a unified set of equations for **sampling from populations of equal- and unequal-sized units, with equal inclusion probabilities** of sampling units within each stratum or **with inclusion probabilities proportional to unit area**, presented in Tyukavina et al. (2025). This set of equations allows sampling unequal-sized map polygons or pixels and pixel blocks from unequal-area grids (e.g., pixel or degree grid in Geographic coordinates) with equal inclusion probabilities and account for unequal area of sample units at the estimation stage. At the same time, this same set of equations is also applicable to the equal-probability sampling of equal-sized units, which is a special case that is covered in Olofsson et al. (2014) and Stehman (2014). Thus, a **unified set of equations** presented in Tyukavina et al. (2025) covers **all one-stage sampling designs** used for global land cover map validation in practice (see Table S1 of Tyukavina et al., 2025), including one-stage cluster sampling.

The estimators in the publications discussed above (Olofsson et al., 2014; Stehman, 2014; Tyukavina et al., 2025) are presented for stratified random sampling, but they can be applied to simple random or systematic sampling as post-stratified estimators, when the strata are incorporated at the analysis stage via the estimator instead of the sampling design stage (Olofsson et al., 2013). When applied to systematic sampling design, equations for standard error estimation are approximations, since systematic sampling does not permit unbiased estimation of variance (Stehman and Czaplewski, 1998), and the true variance is usually overestimated from the sample data when using approximate variance estimators.

For **two-stage cluster sampling**, when a second-stage sample of SSUs is selected from each cluster (e.g., via simple random or systematic sampling, see Figure 3.3.2), accuracy estimators are presented in Zimmerman et al. (2013) and in Chapters 10 and 11 of Cochran (1977). In practice, it is common to use variance estimators of accuracy metrics ignoring the second-stage variance (variance within clusters) to avoid complexity of the two-stage estimators, because the second-stage variance is typically negligible compared to the first-stage variance. This was done, for example, in Potapov et al. (2014), where a one-stage estimator of variance similar to that from Pengra et al. (2015) was used with a two-stage cluster sample.

5.2 Estimating target class area

The literature recommends **estimating the area of a target land cover class** from the **reference classification of a probability sample** (Olofsson et al., 2014), which, as discussed above, should be obtained **independently from the map**, and using **a more accurate labeling method**. Thus, area estimates can be derived from the same reference sample used for map validation with little additional effort.

In this section, we present approaches to sample-based area estimation under the **assumption of negligible error in the reference data**. If the quality of the reference data is low, sample-based estimation becomes problematic (see section 4.2 for discussion related to the quality of reference classification). For methods of assessing the uncertainty in reference data and incorporating it into the variance of the sample-based estimates, please refer to section 4.3.

The recommended sample-based **estimators** of area for specific sampling designs are **unbiased** (see definition in section 2.1.4). Moreover, **sampling theory allows for estimating uncertainty (precision) of the sample-based area estimates**, in the form of variance, standard error, and confidence intervals. The latter is the preferred way of communicating uncertainty, as a confidence interval is easy to interpret (Rice, 2007, p. 217): *"if we were to take many random samples [that could be realized using the sampling design] and form a confidence interval from each one, about 95% [for a 95% confidence level] of these intervals would contain [the population parameter]*".

Often some measure of map uncertainty is reported from classification algorithms, based on the agreement among individual model runs (see sections 2.6 and 7.4). Such model uncertainty measures only show how well the model is able to reproduce the training data, and do not link those map versions with the independently established reference condition. Therefore, uncertainty measures from classification algorithms should not be used as uncertainty indicators for map-based area estimates derived via pixel counting. **A probability sample with reference values obtained independently from the map should be the basis for estimating target class area and its uncertainty.** Note that the map could be used for stratification (see section 3.4) to improve standard errors and this use of the map does not compromise independence of the reference sample from the map. Therefore, the same stratified sample could and should be used both to estimate the reference area of the target class, and to estimate the accuracy of the map that served as a stratifier (and any other map with matching thematic and temporal focus).

Recommended estimators of area

Unbiased area estimators exist for the common probability sampling designs (see section 3.3). Some of these estimators are defined based on the confusion matrix (Olofsson et al., 2013; Stehman, 2013) and initially were referred to as 'bias-adjusted' (Stehman, 2013) or 'error-adjusted' (Olofsson et al., 2013), pointing to the fact that the estimator can be viewed as including the area of map omission error and excluding the area of map commission error (Stehman and Foody, 2019). Such terminology caused confusion among practitioners implementing the methods and certifying agencies formulating requirements for measurement and verification programs, because it creates the impression that the map is somehow ingrained in the area estimates beyond the map's role of providing the strata. In reality, the **sample-based area estimates are always based on the sample reference labels**, and not on the map values. If a map is used to define strata, the strata weights are used in the area estimator formulas, but the actual map values do not impact the area estimates. In fact, the reference area of any land cover class can be estimated in complete absence of a map via simple random or systematic sampling (although using a map to incorporate stratification in the sampling strategy is almost always more efficient, as a smaller sample is needed to achieve a certain precision compared to sampling without a map). Therefore, to maintain clarity that the role of the map is to reduce standard errors, we recommend using direct sample-based area estimators (as opposed to confusion-matrix based), such as indicator functions presented in Stehman (2014) and Tyukavina et al. (2025). We also recommend using the terms 'sample-based area estimator' or 'sample-based area estimate' (see section 2.1.4 for the definitions of 'estimator' and 'estimate') instead of 'bias-adjusted' or 'error-adjusted estimate/estimator'.

Stehman (2014) presents an **estimator of target class area** (equation 3 with y_u defined as equation 14) and its standard error (equations 25 and 26) that is not based on the confusion matrix. This estimator applies to the stratified sampling designs for **sampling from populations of equal-sized units** (e.g., pixel grids with constant pixel size) **with equal inclusion probabilities** of sampling units within each stratum regardless of whether the sampling strata match the map classes. For simple random sampling, these estimators can be viewed as post-stratified, and for systematic sampling, the standard error estimator is an approximation. When sub-unit (e.g., sub-pixel) proportions of the target class are identified in the reference sample classification (e.g., when using high-resolution reference imagery), equation 14 can be modified accordingly to include continuous target class values instead of binary (e.g., proportions such as 0, 0.1, … 0.9, 1.0 instead of 1/0, yes/no).

Tyukavina et al. (2025) provide an extension of the framework presented in Stehman (2014), to accommodate area estimation when **sampling from populations of unequal-sized units** (e.g., map polygons or unequal-area grids, such as pixel or degree

grids in Geographic coordinates) **with equal inclusion probabilities or with inclusion probabilities proportional to unit area**. Tyukavina et al. (2025) also demonstrate the use of ratio estimator to estimate the relative contribution of different subclasses to a target land cover class area (expressed as percentage or proportion) along with its standard error.

The same estimators of target class area from Stehman (2014) and Tyukavina et al. (2025) can be used for **one-stage cluster sampling**, when all secondary sampling units (SSUs) within a cluster are labeled (e.g., all pixels in a block). For **two-stage cluster sampling**, when a second-stage sample of SSUs is selected from each cluster (e.g., via simple random or systematic sampling), area estimators are presented in Jonckheere et al. (2024) and in Chapter 10 of Cochran (1977).

Harmonizing the sample-based area estimates and the map

Once the reference area of the target class is estimated from the reference sample, **a map can be created** or selected from a candidate set of maps (e.g., derived using different input parameters of a machine learning algorithm or using different thresholds to convert continuous algorithm outputs into a categorical land cover map) **such that the pixel counts** in the map **are equal to the sample-based area estimates** for that class **at some aggregated spatial scale** (e.g., the entire map or sub-regions, such as continents). Producing a map with the pixel-based area of land cover class equal to the sample-based estimate does not mean that the errors of commission of omission will be absent in that map, it only means that those errors will be balanced at the selected spatial scale. Examples of studies using this method include mapping wetlands in the Congo (Bwangoy et al., 2010), forest loss in Indonesia (Broich et al., 2011), soybean cover in the United States (Song et al., 2017) and South America (Song et al., 2021), and forest loss due to fire globally (Tyukavina et al., 2022). Examples above are of single-class land cover maps, and therefore selecting a map version with area equal to the sample-based estimate is straightforward. Witjes et al. (2024) propose a general framework for creating a multi-class map with pixel counts for each class equal to independently derived area estimates, which deals with potential class overlaps while selecting the pixels with highest classification probabilities for each land cover class.

Local (i.e., per pixel) map uncertainty measures could also be produced by selecting the versions of the map with pixel counts equal to the sample estimate plus and minus one standard error (or a 95% confidence interval). Incorporating all three map versions (with pixel-based land cover class area equal to the sample-based area estimate, plus and minus the standard error) into one map allows to derive more and less conservative area estimates of the target class via pixel counting (Tyukavina et al., 2022). Such an incorporated map helps reflect sub-regional uncertainties via map pixel counting, because it uses differences between map versions as a proxy for how the uncertainty of

the target class area estimated at the aggregated scale is distributed spatially. However, such an uncertainty layer might be challenging to produce for a multi-class land cover map and therefore is more suitable for assessing local uncertainty of single-class land cover maps or when the multi-class map is converted into a set of single-class maps. For the overview of other methods of local map quality assessment please refer to section 7.4.

Model-assisted estimators of area

Model-assisted estimators utilize available complete coverage (wall-to-wall) auxiliary data related to the target land cover class to reduce the standard error of the target class area estimated from the reference sample (Stehman, 2013, 2009b). An obvious example of auxiliary data for estimating the area of a given land cover class is a wall-to-wall map of that same class. However, spectral information such as a spectral index layer (for estimating the area of a static land cover class), a spectral index anomaly for a certain time period (for estimating the area of a land cover change class), or another continuous metric correlated with the target class could also be used as an auxiliary variable. Although these estimators employ a model to obtain the benefit of enhanced precision, inference is still design-based and the unbiased property of the estimator is not dependent on correct model specification. The more closely the model fits the data, the greater the improvement in precision (Särndal et al., 1992, p.239).

Post-stratification is a model-assisted estimator in relation to simple random and systematic sampling, when the strata are not incorporated into the sampling design but are included at the analysis stage (Stehman, 2009b). Post-stratification might be motivated by the potential benefit of using different strata to reduce standard errors of different target land cover classes (Pickering et al., 2019). For example, a post-stratified estimator can be constructed for estimation of a certain land cover class using a map that contains the target class. Post-stratification yields approximately the same standard error as stratified sampling with proportional allocation of sampling units among the strata (Cochran, 1977, p. 134), because simple random and systematic sampling are equal probability sampling designs, and resulting sample size in post-strata will be approximately proportional to their areas. Post-stratified estimators are effectively the same as the regular stratified estimators of area, as described earlier in this section.

Logistic regression is suitable for model-assisted estimation when the estimation target is binary, e.g., presence or absence of a land cover class (Stehman, 2009b). The logistic regression model can be used with both continuous and binary auxiliary variables. For example, presence or absence of a class can be modeled using a lower resolution binary map of the same class (both target and auxiliary variables are binary) or using a continuous layer of a spectral index values (target variable is binary, auxiliary variable is continuous).

A linear regression estimator (also referred to as simple regression estimator) is suitable for continuous reference and auxiliary data (e.g., when the reference observation is expressed as a proportion of land cover classes) and utilizes the linear relationship between the auxiliary and the target variable (Cochran, 1977, p. 189). For example, Pickering et al. (2021) used population information (auxiliary variable) from 30 m maps (% of target land cover class per 5x5 km block) to increase precision of sample-based estimates of the area of target land cover classes derived from classifying 4 m resolution Planet imagery for each sample block. They observed reduction of standard error of the area estimates from 9.1 to 5.1% for tree cover loss in Peru, and from 9.0 to 3.6% for wheat in Punjab, Pakistan. Such an improvement in precision of estimates, facilitated by using available wall-to-wall auxiliary data, albeit of lower spatial resolution, might be crucial for applications requiring area estimates with a set level of precision. The simple linear regression estimator uses the slope of the linear regression, estimated from the sample data, and the population mean of the auxiliary variable (Cochran, 1977, p. 193). For stratified sampling, two types of regression estimators can be constructed: i) a separate regression estimator with the slope of the linear regression estimated for each stratum separately, and ii) a combined regression estimator with the slope estimated using combined information from the strata (Cochran, 1977, pp. 200-202). In most circumstances the separate regression estimator is preferred.

Difference estimators are similar to regression estimators, but the slope coefficient is fixed instead of being estimated from the sample. If the target and auxiliary variable measure the same quantity (e.g., land cover proportion), a version of the difference estimator, with an intercept of 0 and a slope of 1, is constructed (e.g., Pickering et al., 2019; Potapov et al., 2014). A difference estimator is less flexible than a regression estimator and usually yields a larger standard error but has the benefit of being unbiased for any sample size (Särndal et al., 1992). A regression estimator has a bias of order 1/n (Cochran, 1977), which could become a concern if the sample size is small. Pickering et al. (2019) provide an easy-to-follow application example of difference and regression estimators for simple random sampling (along with appropriate equations), demonstrating the utility of these model-assisted estimators for reducing the standard error of the forest loss area estimates.

6. Sources of reference data

Since 2006 and the publication of the original recommendations for global land cover map accuracy assessment (Strahler et al., 2006), there have been substantial changes in the temporal and spatial resolution of optical satellite imagery sources that can be used to obtain both training and validation data (see section 6.1). In addition, high-resolution remotely sensed data are now available from lidar (also spelled as 'LiDAR' to emphasize the acronym for 'Light Detection and Ranging' - section 6.2) and sensors onboard UAVs (Unoccupied Aerial Vehicles - section 6.3). In 2006, there were very few sources of shared *in situ* reference data; hence, the reuse of existing validation sample sites was a recommendation in Strahler et al. (2006). However, with advances in technologies like mobile phones, cloud storage and processing, there are now many more sources of field-based data, shared through numerous repositories (section 6.4). Although there has been a trend to sample existing land cover maps to increase the amount and coverage of training data (e.g., Radoux et al., 2014; Shetty et al., 2021), this practice is not recommended for validation because of the errors in the existing maps that may introduce biases into the accuracy assessment. Finally, one innovation in reference data collection has been through crowdsourcing, involving both citizens and experts in the visual interpretation of satellite imagery and georeferenced photographs, and buoyed by new technological capabilities and online tools (section 6.5).

In this chapter, 'reference data source' refers to remotely sensed or ground survey data used to determine reference labels for a validation sample (see definitions of 'reference/validation data' and 'reference label/classification' in section 2.1.1). As previously mentioned in section 4.2, the same reference data source should ideally be available for all sample units, and of higher quality than the data used to create the map. For example, higher spatial, temporal and/or spectral resolution satellite imagery can be used as a reference data source to assess the accuracy of the map produced using lower resolution imagery. Or, authoritative ground data collected by experts can be used to validate a map produced using remotely sensed imagery. It is also acceptable to use the same source data to produce and validate the map; in this case the reference labeling method should be more accurate than the mapping method. For example, if the same Landsat time-series are used for both mapping and accuracy assessment, visual interpretation of satellite imagery for a limited number of sample locations is usually employed to derive reference labels, as it is considered a more accurate labeling method compared to automated classification of all pixels in the map during the map production. Please refer to Chapter 4 for further discussion about the reference sample labeling protocol (section 4.1) and quality of reference data (section 4.2).

6.1 Time-series of medium- to very high-resolution optical data

In global land monitoring, optical imagery of varying resolutions is used consistently for training and validation of land cover maps. This includes medium resolution satellite imagery, which is typically 10 to 30 m, to high (<10 m) and very high-resolution (<1 m) satellite and airborne imagery (Skidmore, 2017).

One key source of visually interpreted medium resolution spaceborne reference data is Landsat imagery because of the long time series available (30m resolution from 1982 onward, 60m in 1972-1982) and the open access to the imagery since 2008 (Zhu et al., 2019). For example, the Food and Agriculture's Forest Resources Assessment (FAO-FRA) Remote Sensing Surveys use a visual interpretation of Landsat to complement their statistical assessments (FAO, 2020). FAO's Collect Earth tool uses Landsat imagery in the global assessment of trees, forests and land use in dryland environments (Bastin et al., 2017). Additionally, Li et al. (2017) used Landsat-5 imagery to develop a first all-season training and validation sample for global land cover mapping, for 2015. UMD GLAD Landsat Analysis Ready Data (Potapov et al., 2020) provides a useful reference data source, with cloud-free 16-day and annual Landsat mosaics and the tools for visualizing reference time-series data for each sample unit (available at https://glad.umd.edu/ard/home; an example of a reference data web page is presented in Figure A.4 in the Appendix). TimeSync (https://timesync.forestry.oregonstate.edu/) is another well-established tool for creating time-series reference data for visual interpretation of Landsat imagery (Cohen et al., 2010).

One of the main issues with Landsat satellite imagery, and the use of optical spaceborne data more generally, is the presence of cloud cover (Wulder et al., 2008). Comparatively, the Sentinel-2 constellation currently has a more frequent revisit time (approximately every 5 days, compared with 8 days for Landsat constellation), and thus is more likely to acquire cloud-free imagery for any given location during a given time interval. Additionally, the Sentinel instruments have a higher spatial resolution (up to 10m), which improves the sharpness at which land cover details can be delimited; however, the mission's more recent start (in 2015) means that Sentinel-2 has limited utility for multi-decadal analyses. The Harmonized Landsat-Sentinel 2 (HLS) dataset (Claverie et al., 2018) combines Landsat data since 2013 and Sentinel-2 data since 2015 in a single harmonized dataset at 30 m resolution. The HLS dataset is analysis ready, as it includes atmospheric correction, cloud and cloud-shadow masking, spatial co-registration and common gridding, bidirectional reflectance distribution function normalization and spectral bandpass adjustment. Currently, with Landsat 8 and 9, and Sentinel 2A, B and C in operation, the HLS data revisit frequency is 2-3 days, making HLS data a more useful source of reference data for accuracy assessment (from 2015 onward) than either Landsat or Sentinel-2 data alone, although the spatial resolution of HLS (30 m) is lower

than that of Sentinel-2 (10 m). The projected launches of Sentinel 2D (2028) and the Landsat Next constellation (after 2030) will ensure the continuity of medium-resolution optical reference data sources.

High-resolution spaceborne data (1-10 m, e.g., RapidEye, PlanetScope) provide a middle ground between medium resolution (10-30 m) imagery with higher revisit intervals for each location on the ground, and very high-resolution imagery (< 1 m) with high spatial detail but less frequent revisits. In fact, the PlanetScope data (3-5 m resolution) from a constellation of Dove small satellites, currently have a global daily revisit frequency (higher than medium resolution data), which is revolutionary for map accuracy assessment. The PlanetScope constellation is currently active (imagery available since 2014, but with sparse coverage and lower resolution in earlier years), and the RapidEye constellation (5 m data) was retired in 2020 but has an archive of globally-acquired imagery for 2009-2020. While the entire PlanetScope and RapidEye archive is not freely available, the Norway's International Climate and Forest Initiative (NICFI) has released monthly PlanetScope basemaps for the tropics at 5 m resolution for public use (Sullivan, 2021). NICFI PlanetScope basemaps are available through the Google Earth Engine and via a QGIS plugin, which increases their useability for map accuracy assessment. PlanetScope data has been demonstrated to be an effective source of reference data for estimating land cover change area (Pickering et al., 2021). Another commercial optical satellite constellation, BlackSky, has a capability of targeted data acquisitions in 0.83 - 1.30 m resolution with revisit rate for the same area on the ground of up to 7 times a day. This makes BlackSky data potentially valuable for assessing accuracy of near real-time maps (see section 7.2), although the data are still largely not accessible to the public.

Very high-resolution (VHR) satellite imagery (<1 m) is another reference data source for validating land cover and change maps. Multi-temporal and multispectral VHR data are acquired by passive sensors (e.g., QuickBird, WorldView, GeoEye, Pleiades, SkySat) and are distributed commercially via Maxar, Geo-Airbus and Planet (Schepaschenko et al., 2019). VHR imagery is available for evaluation, research, and product development to governmental agencies such as USDA, NASA, USGS and NOAA. Next, some of this imagery is also openly available via the USGS and ESA's Earth Online portal. VHR data are also commonly distributed as base maps allowing users to seamlessly visualize VHR data across the globe. Examples for openly accessible platforms for base maps are Microsoft Bing Maps, ESRI base maps and Google Earth (Sheppard and Cizek, 2009). Due to this possibility, ease of access and cost efficiency compared to field visits, visual interpretation of VHR satellite imagery has become a standard source for collecting reference data for validation purposes (Schepaschenko et al., 2019; Tarko et al., 2021). Lesiv et al. (2018) analyzed the spatial and temporal distribution of VHR satellite imagery in Google Earth and Bing Maps. The results show an uneven availability globally, with more coverage in certain areas such as the USA,

Europe, and India. They also show that the availability of VHR satellite imagery is currently not adequate for monitoring protected areas and deforestation but is better suited for monitoring changes in cropland or urban areas. Using the Geo-Wiki application, several crowdsourcing campaigns have been run to collect training and reference data, where the datasets are shared in open data repositories (See et al., 2022; section A.1 and Figure A.1.1).

Another source of high and VHR optical imagery that can be used to validate land cover maps is aerial imagery from traditional manned aircraft (see also section 6.3 for the overview of data from unoccupied aerial platforms). Airborne data collection campaigns are typically costly, which limits the feasibility of performing targeted validation campaigns for a specific newly developed land cover map. However, aerial photography previously collected for various land management, cartographic and defense purposes can be used to validate historic land cover maps, if made available to the public. For example, the National Agriculture Imagery Program (NAIP) administered by the U.S. Department of Agriculture, provides 1 m or finer national-scale coverage of aerial imagery for the United States starting from 2003 with 5-year repeat cycle prior to 2009, and no longer than 3 year repeat interval after 2009. Such national-scale datasets are especially useful for land cover validation, as they can provide reference imagery for a probability sample of locations over the entire country. Sub-national and local aerial photography campaigns are less useful for validating national- and global-scale land cover maps but can still provide insights regarding map errors in different environments. Some of the publicly available aerial imagery collections are listed in Table 6.1.

Table 6.1. Examples of online repositories of publicly available aerial optical imagery that can be used for the validation of land cover maps. Links accessed on September 4, 2025.

Repository name	Collection	Temporal and spatial coverage	Spatial resolution	Link
USGS Earth Explorer Catalog	National Agriculture Imagery Program (NAIP) https://doi.org/10.5066/F7QN651G	2003 - present USA national	1m in 2003-2017, 0.3 - 0.6m starting from 2018	https://earthexplorer.usgs.gov/ 'Aerial Imagery' category
	Aerial Photo Mosaics https://doi.org/10.5066/F72805WQ	1937 - 1980 USA national	varies	
	Aerial Photo Single Frames https://doi.org/10.5066/F7610XKM	1937 - present USA national, parts of Latin America, Canada, Europe, Iran, Niger	varies	
	Digital Orthophoto Quadrangle (DOQs) https://doi.org/10.5066/F7125QVD	1987 - 2006 USA national	1 m	
European Data Portal	National aerial photography for the Netherlands	2013 - present, The Netherlands national	7.5 cm, 10cm, 25 cm	https://data.europa.eu/ search 'aerial imagery'
	EU member countries	Various dates and geographic regions	varies	
Geoscience Australia Commonwealth Historical Aerial Photography Collection	Systematically collected imagery, in parallel strips. Additional photographs of the coastline, individual towns and other areas of interest.	1928 - present Australia national, parts of Oceania and Antarctica	varies	https://aerialphotography-geoscience-au.hub.arcgis.com

6.2 Spaceborne and airborne lidar data

Light Detection and Ranging ('LiDAR' or 'lidar') is one of the most recent remote sensing technologies and was in its infancy when the first iteration of this protocol was published (Strahler et al., 2006). Lidar is an active remote sensing technique that emits laser pulses towards the surface and records the returns to build 3-dimensional representations of the Earth's surface (White et al., 2016). Single photon lidars detect the arrival of a single photon from a laser pulse, allowing for high area coverage at a lower cost per data point, while full waveform LiDAR records the entire shape of the reflected laser pulse, providing more detailed information about the target surface, including its texture and shape, at the cost of potentially lower coverage (Mandlburger et al., 2019). Lidar's popularity has increased significantly in the last few years as it can be mounted on a variety of platforms, including unoccupied aerial vehicles (UAVs), planes, and satellites.

Airborne Laser Scanning (ALS) data are a valuable resource for validating land cover maps, as they allow capturing the 3-dimensional properties of the land cover (e.g., vegetation structure, buildings) with high resolution. The resolution of airborne lidar is measured in points/m^2 and is typically between 2 points/m^2 and 8 points/m^2 (White et al., 2016). The cost per unit area to acquire ALS data is high, which limits the spatial extent of the data, typically collected for specific projects at the sub-national level (Hancock et al., 2021). To date, there is not a global dataset of airborne lidar data, which limits the utility of ALS data for validating global land cover maps. ALS datasets often represent a snapshot in time, as they are rarely updated, which diminishes the value of ALS data for the validation of time-series maps.

ALS data collection is often regional, national, or even privately commissioned for commercial use, and therefore locating and/or accessing these datasets for land cover map validation might be a challenge. Several examples of ALS data repositories containing data for different geographic regions are presented below.

- United States: United States Geological Survey (USGS) lidar Collection
- Canada: Library of free lidar products
- Europe: Library of lidar products
- Australia: Lidar for field sites
- Brazil: Lidar surveys over research tropical sites

In contrast, spaceborne lidar (Table 6.2) has the potential for validating global land cover maps, as spaceborne lidar systems have global or near-global coverage, albeit not spatially continuous. A major limitation of spaceborne lidar, compared to optical sensors,

is its narrow ground coverage with each pass due to the need to supply its own illumination (Hancock et al., 2021). For example, passive sensors, such as those onboard Landsat-9, measure a 185 km continuous swath with each pass (Irons et al., 2012). Comparatively, Global Ecosystem Dynamics Investigation (GEDI) mission, which has the widest swath of the spaceborne lidars, has a width of 4.2 km as generated from eight beams, separated by 600 m across track (Dubayah et al., 2020). This is a major limitation for mapping, as it means that to have continuous products, spaceborne lidar data often have to be combined with a continuous dataset, such as Landsat images, so it may be extrapolated (Saarela et al., 2018). For global land cover map validation, the lack of spatial continuity of spaceborne lidar data is less of an issue, as the global sampling strategy of spaceborne lidar systems can be leveraged to derive reference class labels over a global probability reference sample (e.g., use lidar-derived percent woody vegetation within larger reference sample blocks to validate the same variable from a land cover map). However, due to the non-continuous nature of spaceborne lidar acquisitions, their utility might be limited for assessing the accuracy of land cover classes that are small in extent or highly dynamic.

Table 6.2 Publicly available Spaceborne lidar missions. Adapted from Hancock et al. (2021).

Mission	Instrument, type	Years of operation	Ground footprint	Spacing along the ground path	Products	Link to web page
Ice, Cloud and Elevation Satellite (ICESat)	Geoscience Laser Altimeter System (GLAS), full waveform	2003 - 2010	70 m	170 m	Ice sheet mass balance, cloud and aerosols, land topography and vegetation	NASA
ICESat-2	Advanced Topographic Laser Altimeter System (ATLAS), single photon	2018 -	17 m	70 cm	Land ice height, sea ice freeboard, sea ice elevation, land/water elevation data, inland water elevation, ocean elevation	NASA
GEDI	GEDI (Mounted on ISS), full waveform	2018 - (non-continuous)	25 m	60 m	All tropical and temperate rainforests - Canopy cover fraction, leaf area index	NASA and University of Maryland

6.3 Data from UAV

In this section we will address the role of data from unoccupied aerial vehicles (UAVs). Note that here we adopt the use of inclusive language ('unoccupied' vs 'unmanned') following Joyce et al. (2021). UAVs were not mentioned in the previous report on global land cover validation by (Strahler et al., 2006) Strahler et al. (2006). Since then, there have been rapid developments in the field of UAV-technology and routine application of UAVs for various applications (Nex et al., 2022). In this section we describe the potential advantages and drawbacks of using UAVs to validate satellite-derived land use or land cover maps, and highlight possible synergies between satellite and UAV observations, most of which are still under-explored (Alvarez-Vanhard et al., 2021).

In general, there are several features that make UAVs attractive for land cover observations (Berger et al., 2022; Yao et al., 2019). First, the costs of UAVs and the sensors used are low compared to the budgets allocated to satellite missions. Second, UAVs can be directed to specific areas where high spatial resolution images are required, at a lower cost compared to manned aircraft. Third, UAVs can be used to monitor regions at times of specific interest to the user, whereas satellites have fixed orbits and revisit times beyond user control. Furthermore, UAVs can fly below clouds, which often obscure or complicate optical satellite observations.

Despite the benefits mentioned above, there are still some issues that limit the use of UAVs. Although the cost of UAV platforms and sensors is small compared to satellite missions, this can still be a limiting factor for their use, especially in developing countries and compared to satellite-derived data (e.g., Landsat missions, Sentinel), that individuals can access free of charge through imagery services and cloud-computing geospatial platforms (e.g., Google Earth Engine, USGS EarthExplorer). In other words, the higher cost of satellite missions is often shared by the society, whereas the lower UAV platform and sensor costs are often borne by the specific research projects. In addition to equipment cost, UAV reference data collection also includes travel time and costs for the UAV operator, who needs to be within a few kilometers from a data collection site due to the limited range of UAVs. In this respect UAV reference data collection is similar to traditional ground surveys (section 6.4). Further, technical training is required to operate the UAV and to process data into a usable format for further analysis (Yao et al., 2019), whereas many satellite-based products are provided to the users in analysis-ready formats (see section 6.1).

Besides these practical and technological concerns, the application of UAVs is further limited by government regulations (Stöcker et al., 2017). These regulations can vary between countries and even at sub-national level. For safety reasons, flying may be prohibited in areas in proximity to airports and industrial or military complexes, or above

certain maximum altitude. These legal restrictions can, for some regions, limit the possible range of operation of UAVs and thereby impair their potential to observe the land cover.

Given the low cost, the ability for directed monitoring of areas at user-specified revisit times, and for capturing cloud-free imagery, UAVs can provide valuable regional information for validation of satellite-based land cover datasets. UAVs can serve as a 'last km' solution for *in-situ* data collection, i.e. collecting reference data for validation sites located away from the roads or in rugged terrain and not easily accessible by ground transport or on foot. UAVs are particularly valuable for highly heterogeneous areas, including urban areas that contain irregularly shaped patches of vegetation (e.g., Al-Najjar et al., 2019; Natesan et al., 2018) and island ecosystems (Laso et al., 2020). As UAVs yield the potential for high temporal resolution of acquisitions, they might be useful in monitoring landscapes that have been exposed to rapid changes, such as fires (Lazzeri et al., 2021) and urbanization (Jumaat et al., 2018). Detailed regional maps produced using UAV data could be then incorporated into the accuracy assessment of national, regional and global land cover maps.

The field of UAV land monitoring is still very much in progress. Looking forward, we expect that further improvements in UAV technology (including batteries, sensors, and software) will increase the synergies between UAVs and satellites (e.g., by allowing UAVs to monitor larger areas because of longer flight time, or to record data in more spectral bands). Moreover, the development of open-source software for UAV observation and processing (De Luca et al., 2019; Horning et al., 2020), coordinated data collection strategies, and open data sharing policies (Dwivedi et al., 2022; Koren et al., 2022) can further accelerate the uptake of UAV data in the field of land cover classification and map validation.

6.4 Ground surveys

The current section focuses on the *in situ* data obtained via field visits that can be used to assess accuracy of land cover maps. In any remote sensing application, the data captured via 'boots-on-the-ground' is essential for product validation. Concomitantly to the increasing volume of remotely sensed data, there is an attendant need to improve the availability of *in situ* data, in order to improve the reliability of remotely sensed data products and algorithms. To date, the sharing and standardization of *in situ* land cover data are very limited among the research community. One of many reasons is the lack of cyberinfrastructure for researchers to share *in situ* land cover data (Szantoi et al., 2020). Other reasons related to data privacy, ownership and intellectual property rights might preclude sharing *in situ* data such as forest plots (de Lima et al., 2022) and agricultural data (Ellixson and Griffin, 2016), both of which could be useful for land cover and land use map validation.

In situ data can be divided into categories linked to their provenance: 1) field surveys and sampling, 2) administrative data and 3) crowdsourcing. Examples of each type of *in situ* data are provided below.

A traditional field survey means going into the field and collecting information for a specific project (e.g., validating a land cover map). Field surveys can collect data on the entire population (e.g., a census) or collect a subset containing a sufficient number of sample locations (e.g., random or systematic). When considering data collection for training or validation of satellite-derived products, field sampling can be undertaken to collect data on representative traits and land covers related to satellite imagery by date and/or other categories (e.g., species; functional groups; nutrient and water availability) or to target land cover changes. Examples of collecting field data for a specific project include national- and regional-scale crop type map validation and area estimation (Khan et al., 2016; Song et al., 2021, 2017) and attributing drivers of mapped forest loss events (Krylov et al., 2018). The data collected through field sampling are deposited in existing repositories (see Table 6.4) or published along with peer-reviewed publications, together with a description of the sampling design and collection methods. The types of the data collected in the field for specific projects may vary widely and therefore, and therefore we do not aim to develop a standardized metadata for sharing the field survey data in the current document. We provide some key requirements for the metadata reporting in section 2.2, and in Tables 3.1 and 4.1. Some communities within the broader land cover mapping field, however, have formulated their own guidelines for field data collection, e.g., guidelines for cropland and crop type field data collection (Group on Earth Observations, 2018) within the Joint Experiment for Crop Assessment and Monitoring (JECAM).

Administrative surveys performed for statistical purposes could also be used for land cover validation, in addition to the field survey data collected for a specific project or product validation. For example, the European Union (EU) Land Use/Cover Area frame Survey (LUCAS) includes over 1.3 million observations in over 650,000 unique locations, from 2006 and 2018, for 106 variables including, for example, land use and land cover, crop residues, and soil samples (d'Andrimont et al., 2020). This is the most comprehensive *in situ* dataset on land cover and use in the EU. LUCAS has been used to validate several land cover maps (e.g., Karydas et al., 2015; Venter et al., 2022; Verhegghen et al., 2021) while initially designed for land cover statistics only. For more details on the LUCAS dataset and moving towards standardized reference datasets please refer to section 7.3. A similar land cover and land use survey is being conducted over Africa with 20,000 sample units (https://www.soils4africa-h2020.eu/the-project). Another example of administrative surveys that are used as reference data for map validation is the US Department of Agriculture's National Agricultural Statistics Service (USDA NASS) farmer surveys. NASS surveys are used as inputs for the USDA's annual

Cropland Data Layer, which is a national crop type map (Johnson, 2013). The EuroCrops dataset provides reference cropland polygons reported by countries of the European Union (https://www.eurocrops.tum.de/index.html), and is included the WorldCereal reference module (https://esa-worldcereal.org/en/about/reference-data), which is more exhaustive and delivers global collection of cropland reference data for cropland mapping (Boogaard et al., 2023).

Crowdsourcing and citizen science approaches to collect and share *in situ* data are relatively new. Some of these approaches aim to collect land cover and land use data as their primary purpose such as the Geo-Wiki project at the International Institute of Applied System Analysis (IIASA) in Austria (http://www.geo-wiki.org, Fritz et al., 2012). There are other initiatives that collect data for other purposes but might be useful for training and validation of land cover and land cover change maps. Examples include the NASA-funded Global Observer program (https://observer.globe.gov/),, the Global Geo-Referenced Field Photo Library at the University of Oklahoma (https://www.ceom.ou.edu/photos/, Xiao et al., 2011), the Geograph project (https://www.geograph.org/) and the Degree of Confluence project (https://www.confluence.org/, Iwao et al., 2006). All these efforts utilize smartphone apps for capturing photographs in the field, and/or websites where users can visualize and share field photographs. To date, no effort has been made to combine the available field photographs across the global websites and other initiatives, to develop a standard training and validation dataset. Field photographs from the aforementioned initiatives and street view photographs from common web map platforms (e.g., Google and Bing maps) can be interpreted and classified into various land cover types by visual interpretation or by deep learning algorithms and subsequently used as training data for mapping algorithms (d'Andrimont et al., 2022; Xu et al., 2017). However, automated land cover labels from photographs should be used with caution for map validation purposes: careful quantification of errors and biases of such automated photograph labeling should be performed and compared with those of manual photograph labeling by humans to identify whether algorithmically derived land cover labels can serve as a reliable reference data source.

Field data are made available to the Earth observation community in either specialized or diverse repositories. The specialized archives are frequently associated with a specific organization or a group of organizations which requires its members to deposit their data there. Comparatively, diverse repositories accept a range of categories of field data from any member of the supporting organization. In both types of repositories, the data are referenced and accessed through, for instance, peer-reviewed curated datasets associated with articles, online databases, or data archives. Below, we provide a list of common Earth observation field data repositories (see Table 6.4, which was compiled in December 2022). The use of peer-reviewed repositories is recommended, to

ensure the quality of the data and methods of collection. Furthermore, there is a strong need for a centralized search engine to link the peer-reviewed field datasets from various repositories, making them discoverable by interested professionals from a specific field. Google's new pro data set search facility (https://datasetsearch.research.google.com/) is a prototype of such a search engine.

Field data can be published in a peer-reviewed article including a description of the sampling methods of collection, targeted data applications and future uses. Examples of multidisciplinary journals that provide access to the research data and the associated research publications include the Elsevier journal *Data in Brief,* Nature's *Scientific Data*, MDPI's *Data* journal and Copernicus's *Earth System Science Data* journal. Since remote sensing is an interdisciplinary field, we encourage the use of collections such as the Registry of Research Data Repositories (Re3data.org) and FAIRsharing.org to find the appropriate repositories for your field data.

Table 6.4. Examples of online repositories of field data for training and validation of remote sensing products. Specific repositories maintain data collections on specific topics and/or are limited to the host institution(s). Diverse repositories maintain multidisciplinary collections of open-source data, accepted after peer review and organized by discipline or application community. Links accessed on September 4, 2025.

Repository name	Type	Link
4TU.ResearchData	Specific to data from Delft University of Technology, Eindhoven University of Technology, Twente University and Wageningen University	https://data.4tu.nl/info/en/
EnviroNet	Specific to University of Alberta	https://www.enviro-net.org/Default.aspx
Environmental Data Initiative (EDI)	Specific to NSF	https://edirepository.org/
Figshare	Diverse	https://figshare.com/
FAIRsharing	Diverse, descriptions of community standards and databases	https://fairsharing.org/
Global Biodiversity Information Facility (GBIF)	Diverse	https://www.gbif.org/what-is-gbif
Global Index of Vegetation-Plot Databases (GIVD)	Diverse, plant communities	https://www.givd.info/info_organisation.xhtml
Harvard Dataverse	Specific to Harvard	https://dataverse.harvard.edu/
HydroShare (CUAHSI)	Diverse	https://www.hydroshare.org/
ICOS: Integrated Carbon	Specific to Integrated Carbon	https://www.icos-cp.eu/data-

Observation System (Europe)	Observation System (Europe)	services/about-data-portal
ImagineS : Implementation of Multi-scale Agricultural Indicators Exploiting Sentinels	Specific - CGLS	http://fp7-imagines.eu/
Integrated Taxonomic Information System	Specific to NSF, Smithsonian and USGS	https://www.itis.gov/
Interdisciplinary Earth Data Alliance	Specific - Geosciences, NSF	https://www.iedadata.org/
KNB: The Knowledge Network for Biocomplexity	Diverse	https://knb.ecoinformatics.org/
Land Processes Distributed Active Archive Data Center (LP DAAC)	Diverse	https://lpdaac.usgs.gov/
Mendeley Data	Diverse	https://data.mendeley.com/
National Tibetan Plateau/Third Pole Environment Data Center	Specific to the Tibetan plateau	https://data.tpdc.ac.cn/en/
NEON: National Ecological Observatory Network (US Only)	Specific to NEON	https://www.neonscience.org/
NOAA National Centers for Environmental Information	Specific to NOAA	https://www.ncei.noaa.gov/
Oak Ridge National Laboratory	Diverse	https://www.ornl.gov/
Open Science Framework	Diverse	https://osf.io/
data.europa.eu - The official portal for European data	Diverse	https://data.europa.eu/en
Joint Research Centre Data Catalogue	Diverse	https://data.jrc.ec.europa.eu/
Registry of Research Data Repositories	Diverse, Global registry of research data repositories	https://www.re3data.org
Science Data Bank	Diverse	https://www.scidb.cn/en
TRY Database	Diverse, global database of curated plant traits	https://www.try-db.org/TryWeb/About.php
U.S. Department of Energy's (DOE) Environmental System Science Data Infrastructure	Specific to DOE	https://data.ess-dive.lbl.gov/data
Zenodo	Specific to NASA's Transform to OPen Science (TOPS)	https://zenodo.org/

6.5 Expert-based methods of reference data collection vs. crowdsourcing

Expert-based methods of reference data collection involve the use of scientists or professionals trained in remote sensing, spatial sciences, visual interpretation and/or field-based survey methods to collect reference data for accuracy assessment of land cover/land use products. There are many examples of where this approach has been used in the past, for instance, to validate global land cover maps such as GlobCover, and land cover time series such as ESA-CCI and The Copernicus Global Land Service – Dynamic Land Cover 100m (CGLS-LC100) (Bicheron et al., 2008; Bontemps et al., 2011; Defourny et al., 2017; Tsendbazar et al., 2020). The reference data collected using expert-based approaches are of relatively high quality but not error free (section 4.2).

To date, several studies have begun to address errors in reference data derived from expert interpretations. For example, Foody (2010) simulated different types of reference data errors and found that even a small amount of reference data error can result in large errors in the resulting accuracy metrics of land cover change. Foody (2010) provides methods to reduce or remove the effects of reference data error, from simple algebraic methods to fitting models such as latent class analysis. McRoberts et al. (2018) simulated the effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions when simple majority interpretation of multiple experts evaluating the entire sample is used as final reference labels. They found that bias increased with greater inequality in strata sizes, smaller map and interpreter accuracies, fewer interpreters and greater correlations among interpreters, of which only the number of interpreters can be controlled if interpreters are working independently. McRoberts et al. (2018) also found that failure to account for interpreter error produced stratified standard errors that under-estimated actual standard errors of the estimated quantities (areas of land cover classes or map accuracy metrics). A greater number of interpreters mitigates the effects of interpreter error on estimates, and a hybrid variance estimator presented in McRoberts et al. (2018) allows to account for the effects on standard errors when the entire sample is interpreted multiple times (once by each individual interpreter). The method presented in Stehman et al. (2022) allows interpreter variability to be incorporated into estimation of the total variance when only a random subsample of a reference sample is interpreted by multiple experts, which provides sample interpretation cost savings compared to re-interpreting the entire sample by multiple experts. Even greater cost savings for incorporating interpreter variability might be achieved by using interpenetrating subsampling (Xing and Stehman, 2024), in which the full sample is partitioned into several nonoverlapping subsamples each evaluated by only one interpreter, without repeat interpretations. Note that these studies (McRoberts et al., 2018; Stehman et al., 2022; Xing and Stehman, 2024) discuss estimation of the land cover class

area, but the same general methodology and findings apply to the sample-based map accuracy estimation as well.

As an alternative to majority interpretations from a large number of independent interpreters, McRoberts et al. (2018) suggests using the consensus interpretation approach, in which interpreters discuss the specific sample units to reach consensus on reference sample labels. Usually, the initial round of sample labeling in the consensus approach is performed by each expert independently, with each sample unit being interpreted by at least two experts. Then, the sample units that have disagreements in the initial labels from different experts are discussed collectively to reach consensus regarding the final reference label. This allows focusing interpretation effort on the uncertain cases, e.g., acquiring additional reference data sources for those uncertain cases, consulting regional experts or doing in-depth information searches online. The consensus approach has been adopted in validation of several global land cover maps (Bassine et al., 2020; Hansen et al., 2022, 2013; Potapov et al., 2022a; Tyukavina et al., 2022), with the reasoning that the consensus approach allows the interpreter uncertainty in the reference data to be minimized instead of incorporating often significant among-interpreter variance from their initial independent sample interpretations into the final variance estimates. For further discussion related to the quality of reference data and interpreter disagreement please refer to sections 4.2 and 4.3.

In contrast to the use of experts, reference data can also be collected using crowdsourcing, which is defined as the outsourcing of tasks that would otherwise not be possible to perform with existing resources to a large number of people, 'the crowd' (Howe, 2006). Other terms for involving 'the crowd' in data collection include 'citizen science' (Bonney et al., 2009) and 'volunteered geographic information' (Goodchild, 2007), among many others (See et al., 2016). The commonalities amongst these different terms include (i) the use of a crowd or a set of volunteers, who are generally not experts in visual interpretation or ground-based data collection; and (ii) the increasing use of technology to facilitate the data collection process, e.g., using online applications such as Collect Earth (Saah et al., 2019) and Geo-Wiki (Fritz et al., 2012) and mobile apps (Bayas et al., 2020). Broadly, crowdsourcing draws upon an engaged public to provide environmental observations. In measuring the health of species populations, Johnson and Sumpter (2016) showcased the immense utility that anglers and birders have provided to community monitoring in the United Kingdom. Further, this sentiment is echoed in cellphone applications like iNaturalist, where citizens can learn about their surrounding environment, while providing essential reference data for global plant and animal species distributions (Aristeidou et al., 2021). Amongst the advantages of a crowdsourced approach to reference data collection are the potentially large amounts of reference data that can be collected. For example, See et al. (2022) summarizes the amount of data that has been collected across several crowdsourcing campaigns, which can be on the order

of hundreds of thousands. The German Aerospace Center (DLR) collaborated with Google to collect a very large reference dataset (900K sample sites) to validate remotely sensed built-up products (Marconcini et al., 2020). Table 6.5 provides advantages and disadvantages of both approaches, which can be used to guide the choice of which method to use. The amount of reference data collected may be one of the key factors that governs this choice.

Finally, it is possible to have a hybrid approach in which 'the crowd' are experts (and/or students) from different fields, who could be trained through workshops to aid in reference data collection. This way, the amount of reference data collected can still be very large but investments in training materials and continued support are still required. However, the data quality should be higher from such a hybrid approach, compared to a pure crowdsourced approach. Lesiv et al. (2022) employed this approach to collect more than 130K reference data points on forest management, inviting forest experts to a workshop, and training them in visual interpretation. Waldner et al. (2019) discuss the specifics of conflation of expert-based and crowdsourced reference data and provide recommendations for collecting hybrid reference datasets. Stehman et al. (2018) demonstrate how expert and crowdsourced data can be usefully combined even if the crowdsourced data have not been acquired using a probability sample. The approach combining crowdsourced labeling of the entire reference sample and expert consensus-based labeling of the sample subset was developed for assessing the accuracy of participatory burned area mapping (Glushkov et al., 2021).

Table 6.5 Advantages and disadvantages of expert-based vs crowdsourced reference data collection

Reference data collection	Expert-based	Crowdsourced
Advantages	Reference data are of a higher quality than crowdsourcing-based data collection but not error free. Reference data for validation are based on a probability sampling design. Required interpreter training is minimal; training materials and tools could be more technical, no need to adapt them for non-specialists. Interpreter error can be minimized through consensus approach or incorporated into the final estimate.	Very large reference datasets can be collected, allowing map accuracy assessment at finer geographic scales. Typically, more repeated interpretations of the same sample unit compared to expert-based methods, which in the absence of unidirectional bias/high correlation among the crowd-interpreters can reduce the bias in the accuracy and area estimates (McRoberts et al., 2018). The models can be used to assess data quality and to filter the data. Greater potential to incorporate local knowledge through a large volunteer base. Interactive tools can be created to reinforce/reward correct interpretations.
Disadvantages	The amount of reference data that can be collected using highly trained experts is limited thus often not allowing for finer-scale accuracy assessment (e.g., estimating sub-continental or national accuracy metrics in global assessments). It may not be possible to incorporate local knowledge if only a few experts perform global or large regional assessments.	Data quality is generally lower than expert-based data collection but there are also examples of where crowdsourced data are better than expert/authoritative data. Local crowd-interpreters are not necessarily familiar with the variety of land covers in their region and their representation in the satellite imagery. Requires time to train volunteers, to create training materials, to provide continuous feedback, to collect gold standard reference datasets for assessing crowd quality and to create mechanisms of assessing the quality of individual interpretations to remove the outliers. Field-based data collection may be opportunistic and hence can be biased towards urban/easy to reach areas and more suitable for model training/testing rather than validation, which requires a probability sampling design. However, directed field-based data collection may result in compliance with a probability sampling design.

7. Challenges and future directions

Global land cover and change maps are used in a wide range of studies to understand Earth system functioning and change (Foley et al., 2005; Song, 2023; Townshend et al., 1991). In other practical applications, such as crop area reporting or estimation of higher-level variables for carbon accounting, the area estimates derived from land cover and change maps are the end results (Friedlingstein et al., 2023). Accuracy information of global land cover data is thus critically important for downstream users. Despite the well justified scientific value of accuracy assessment and the existence of community guidelines (Strahler et al., 2006), many existing global data products still lack validation, are at low validation stages, or lack reproducibility and transparency in reporting of methodology or reference data (Table 1.4).

The availability of medium-resolution satellite data in easy-to-use format (e.g., Analysis Ready Data), and substantial advances in machine learning and computing, have considerably reduced the technical barriers of producing global land cover maps. The developed technological infrastructure further enables operational updates of global land cover maps at annual or finer frequencies. Even more so, advanced algorithms for near-real-time (NRT) change detection and time-series analysis are becoming mature for NRT disturbance monitoring in an operational mode. Object-based image analysis is also an increasingly popular approach for land cover mapping with high-resolution remote sensing data and deep learning algorithms. Moreover, global land cover mapping is moving toward generating essential, desirable and aspirational global products along the land cover and land use hierarchy that are directly relevant to sustainable land management and societal benefits such as crop types, forest types, or urban structure (Hansen et al., 2022; Radeloff et al., 2024).

Advances in land cover mapping bring new challenges and opportunities for validation. Assessing the accuracy of object-based maps introduces new issues not present in traditional pixel-based accuracy assessment with regards to the sampling design, response design and analysis (Stehman and Foody, 2019). For operational updates of new land cover products, the sampling design, reference labels and analysis need to be updated accordingly. Strategies for updating the stratification and periodic revisiting of the reference datasets are required to reach Stage 4 validation (see more details in section 7.1). The dynamic nature of NRT monitoring systems poses great challenges for validation as compared to traditional validation of historical land cover change. Validation of NRT systems should be designed for assessing the timeliness of land change events in addition to assessing the spatial accuracy (section 7.2).

Technological advancements have substantially expanded the various means for geospatial data collection, ranging from frequent revisit of very-high-resolution satellites to unoccupied aerial vehicles (UAVs) and mobile platforms (see Chapter 6). Although

data collected through these novel means may or may not follow a probability sampling design, they can be integrated with a probability sample to enhance accuracy assessment (Fonte et al., 2015; Stehman et al., 2018). Coordinated efforts are emerging that are helping to move toward standardized reference data collection, for example, the European Union Land Use/Cover Area frame Survey (LUCAS, section 7.3). In addition to quantifying thematic accuracy for the entire map or its subsets via overall (e.g., overall accuracy) and class-specific (e.g., user's and producer's accuracy) accuracy metrics, future efforts should also consider local map quality. Local map quality metrics can be obtained by using the uncertainty outputs of machine learning algorithms or by predicting the probability of correct classification of a map unit through spatial interpolation of agreement with an independent sample of reference labels (section 7.4).

7.1 Operational validation updates

Remote sensing-based applications are evolving toward continuous monitoring instead of mapping the Earth's surface for limited point(s) in time (Woodcock et al., 2020). Accordingly, several activities provide operational and continuous land cover products. For continuous land cover monitoring, temporal consistency in land cover characterization over time is important (Bontemps et al., 2011) and is highlighted as a key requirement for monitoring land cover as an essential climate variable (ECV) for Global Climate Observing Systems (GCOS) (GCOS, 2011, section 1.5).

With the land product validation guidelines developed by the CEOS LPV subgroup (section 1.2), and other land cover validation community efforts, such as GOFC-GOLD and their best practice guidelines, most of the recently published global land cover maps are now validated to the CEOS LPV stage 3 (see Table 1.4) using statistically rigorous accuracy assessment methods (Olofsson et al., 2014; Strahler et al., 2006; Xu et al., 2020).

For operational land cover mapping, updated products are also subjected to updated validation. The CEOS LPV stage 4 validation guidelines recommend a systematic updating of validation results for each new release or time series expansion of land products (section 1.2). Therefore, to support users' confidence in the continued use of land cover products, operational land monitoring efforts need to expand their product validation into operational validation by regularly updating product uncertainty and consistency information. To update the product validation with a new release, the validation design of the initial/base product can be adapted provided that the validation sampling design is flexible for changes in the map introduced by map updates and the reference labels in the validation dataset are up to date. The sections below focus on these two aspects of operational validation updates.

Sampling design considerations

The key aspect of operational land cover mapping is to provide a regular update on the product by mapping or incorporating changes that occurred within the update interval (e.g., annually or every 3-6 years). For operational validation, the sampling design needs to be flexible or adaptable to account for such changes without compromising the statistical rigor (see Chapter 3 for the general principles of sampling design). In the case of an operational national monitoring program (Pengra et al., 2020), simple random sampling was initially implemented to allow collection of reference data prior to completion of a map. This design allowed for a later sample augmentation targeting particular classes or change areas based on map classes, once the map was available, and the combined sample was used to estimate accuracy and area (Stehman et al., 2021). Stratified random sampling that uses different stratification than the mapped product was adapted in the operational validation of global land cover maps (Tsendbazar et al., 2021). This design also enables augmenting a validation sample to address specific regions such as change areas or classes of interest (Stehman et al., 2012) and parallel data collection.

For validating the new release of a land cover product, a stratified sampling design based on the 'base' land cover maps is also applicable. However, due to the difference in the strata used for the initial sample and map classes of the new release, accuracy estimation needs to be adjusted. In this case, instead of the commonly used stratified sampling estimators by Card (1982) and Olofsson et al. (2014), the estimators from Stehman (2014) and Tyukavina et al. (2025) should be used because they are not based on the assumption of map classes matching sampling strata. Other sampling schemes such as systematic random sampling and cluster sampling could also be investigated in an operational validation context without compromising statistical rigor and timeliness.

If there is an interest in gaining insight on land cover changes, the operational validation framework can be expanded to suit this need. A main requirement for land cover change validation is an adequate sample size of land cover change areas to ensure estimates of user's accuracy are sufficiently precise. To satisfy this, the framework proposed by Tsendbazar et al. (2021) for operational validation of annual land cover maps can be adapted (Figure 7.1). Here, to increase sample size in change areas, the initial sample for a certain reference year (T0) is augmented for land cover change areas in a later period or year (T1) (Figure 7.1-3). Subsequently, the original stratification is modified by adding the new change stratum and recalculating the sample inclusion probability for all sample sites (Figure 7.1-4). The augmented sample sites in T1 can be considered temporary sample sites (Figure 7.1-3) and are not used in validating further releases (e.g., T2), to simplify the stratification. For further updates, the up-to-date original validation dataset ('Original sample sites' and 'Revisited sample sites' combined, Figure 7.1-4), without the added sites ('Additional change area sample sites', Figure 7.1-4) is considered the starting point, and the same procedure is followed to update the dataset further.

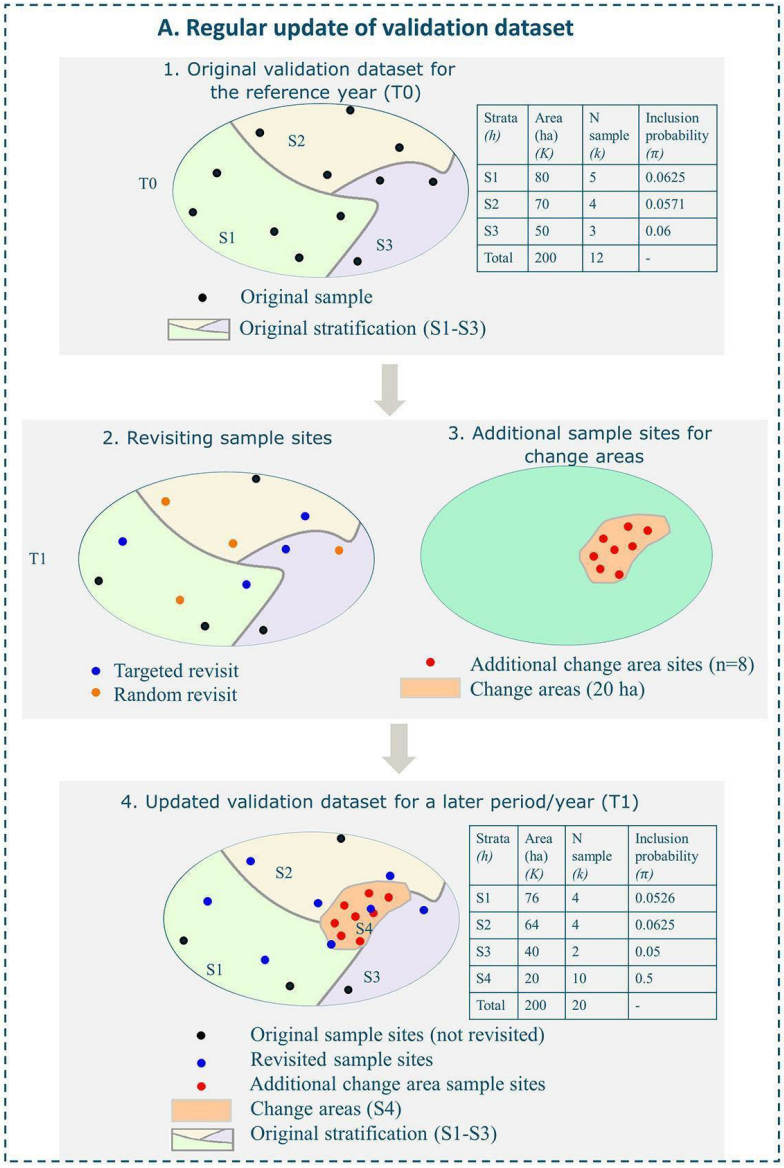

Figure 7.1 A framework for operational validation for land cover monitoring. Figure from Tsendbazar et al. (2021).

Updating reference labels in validation data

Map validation assesses the accuracy of a mapped product using a reference dataset that is of higher quality than the map being evaluated (Olofsson et al., 2014). The validation datasets are often created through visual interpretation of very high-resolution images (Tarko et al., 2021). Please refer to Chapter 4 for more information about recommended response design (protocol of assigning reference labels).

Since the manual revisiting and interpretation efforts are very costly and time-consuming, efficient ways of keeping the reference labels up-to-date are required for operational validation. Partially revisiting the validation dataset consisting of a targeted

revisit and a random revisit was proposed in the operational validation of annual global land cover maps (Tsendbazar et al., 2021) (Figure 7.1). First, limited resources are available to revisit (by means of visual interpretation) sample sites that have a high possibility of land cover change occurrence since the reference year of the initial validation data (targeted revisits, Figure 7.1). A change detection algorithm can be conducted on time series satellite data to identify sample sites that have a high possibility of land cover change. Change detection algorithms such as Breaks For Additive Season and Trend (BFAST) and Continuous Change Detection and Classification (CCDC) can be run at validation sites (Verbesselt et al., 2010; Zhu and Woodcock, 2014). Second, depending on the available resources, a random subset of the validation dataset is rechecked for possible land cover changes. Such a random revisit is particularly useful to assess whether change detection algorithms are omitting any occurrence of the land cover change. Considering the efforts and time required for manual revisit and interpretation of validation sites, a partial and targeted revisit instead of a full revisit can be an efficient way to reduce delays between map accuracy estimation and map update release.

There is still a possibility of change detection algorithms missing sample sites with land cover change and an un-revisited site has seen a change in land cover. This can have an impact on the quality of the validation data. Such issues could be alleviated by improving the quality of the land cover change detection algorithm. Other developments in change detection (Asokan and Anitha, 2019) and high-resolution satellite data such as Sentinel 1 and 2 and Planet could be investigated to identify validation sites that have a high possibility of change occurrence and visually confirm. At the same time, possible errors in the validation data could be taken into account when assessing land cover map accuracies (Stehman and Foody, 2019) (see sections 4.2 and 4.3). Regardless of the efficiency of the change detection algorithm or reference data error, a full revisit of the validation dataset after a certain period (e.g., 5 years) is still recommended to maintain the quality of the dataset, particularly with the increased availability of very high-resolution images over time. This highlights the importance of maintaining validation data including regular rechecking and maintaining validation data collection interface (e.g., those based on Geo-Wiki) for operational validation.

7.2 Assessing accuracy of near real-time maps

Advancements in satellite remote sensing and cloud computing have enabled near real-time monitoring of land changes on local to global scale, providing valuable insights on the status of human-environmental systems and facilitating timely action in time-sensitive situations such as responding to illegal logging (Hansen et al., 2016; Nagatani et al., 2018; Reiche et al., 2021; Shimabukuro et al., 2007) or active fires (Giglio et al., 2016; Justice et al., 2002; Schroeder et al., 2014). Early warning systems can also identify

ongoing or potential near-future threats (Becker-Reshef et al., 2019), which helps to mitigate the environmental and human impacts of emergent events such as invasive species outbreaks or natural disasters. While methods for the assessment of near real-time monitoring systems is a topic of ongoing research, their development and use is expanding rapidly. Therefore, there is a need for guidance on how to use and evaluate these systems effectively. This section aims to provide a starting point by outlining key concepts, definitions (Table 7.2), and initial guidelines (see below).

Table 7.2 Key terminology related to satellite-based near real-time monitoring of land changes

Near real-time monitoring system	A set of procedures whose main purpose is to identify particular land changes continuously and as quickly as possible. The specific definition of 'near real-time' will vary based on the use case and can range from seconds to months.
Early warning system	A set of procedures whose main purpose is to identify preliminary indications of potential land changes.
Alert	A detection/early warning of a land change, typically delivered in a spatial format such as a pixel or polygon on a map or the geographic coordinates.
Time Lag	The time interval between a land change event and its identification by an algorithm due to both observation lag (e.g., sensor temporal resolution, clouds) and algorithm lag (e.g., data latency, algorithm constraints).
Timeliness	A general term summarizing the time lag at which a monitoring system identifies land changes.

Traditional remote sensing approaches for assessing historical land cover or land use change often use similar methods and data as systems that operate in near real-time. Both types of analysis often apply change detection techniques to multi-temporal satellite data to map or quantify the area of land change. Although extensive research on historical land change monitoring has resulted in guidelines for proper use and evaluation of land change maps (GFOI, 2020; Olofsson et al., 2014), there are critical differences that make these community-accepted standards incomplete in the near real-time context.

Unlike historical change analysis, the timeliness of a near real-time monitoring system can be as important as its spatial accuracy (Bullock et al., 2022; Reiche et al., 2018). For example, one application of near real-time forest monitoring is to identify illegal logging operations that often occur over a few days. In these cases, the speed of detection can be more relevant to law enforcement interdiction than spatial accuracy. However, it can often take weeks to months after the event for an alert to be created due to various environmental and technical factors (e.g., clouds, sensor revisit times, or algorithm parameters) that introduce lag into the system (Figure 7.2). It is therefore critical to consider both spatial accuracy and timeliness when evaluating a near real-time monitoring system.

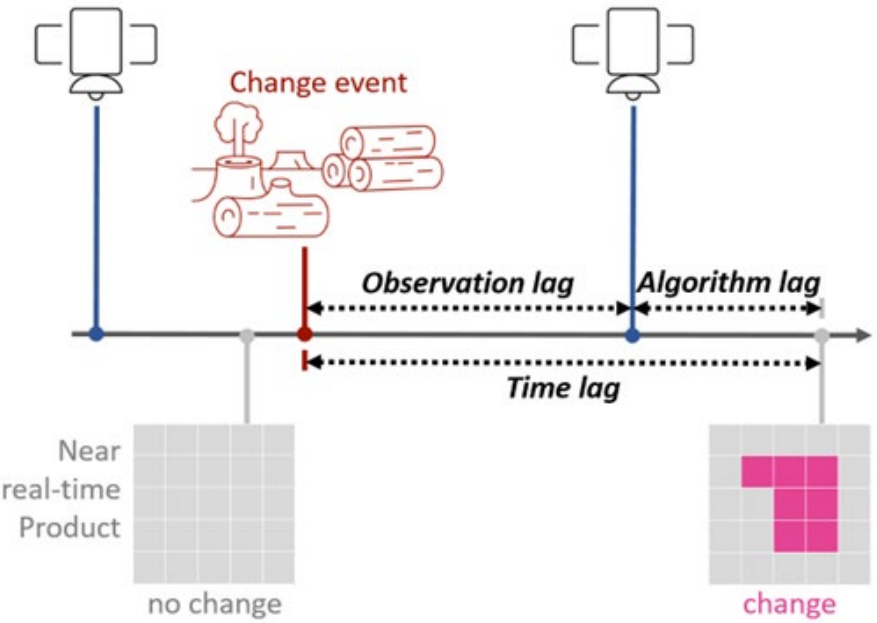

Figure 7.2 The time lag between a land change event and an alert is determined by the latency between the event and the time it is observed by a satellite (observation lag) and the delay of satellite data delivery and configuration of the change detection algorithm (algorithm lag). Some systems may require multiple observations to identify changes. Figure by Johannes Reiche, Amy Pickens and Eric Bullock.

In the previous example, a change alert may be useful if it simply directs the enforcement agency to the general area of a logging event, even if a particular pixel is mapped incorrectly. Land change events often occur in spatially continuous patches within a limited time interval, suggesting that the use of events rather than pixels for evaluating system performance may be preferable in the near real-time context (Bullock et al., 2022; Tang et al., 2019).

Near real-time systems are also dynamic, meaning that change alerts (or alert confidence) may vary as more data becomes available, and the map accuracy at a specific point in time may not reflect its overall performance. The dynamic nature and other factors described above highlight the unique considerations for evaluating the performance of a near-real time monitoring system.

Initial guidelines on assessment design

While event-based validation concepts for assessing near real-time maps have started to evolve, pixel-based approaches are more common and more straightforward to apply. This initial guidance on assessment designs presents an approach using pixel-based spatial units and presents only one solution based on recent studies. Regardless of the approach taken, the timeliness should be assessed together with the spatial accuracy of the maps. A clear definition of what constitutes an event that should be

captured by the change class is required. Further, given the typical evolving nature of change events, it must be clearly defined at what stage of the process it should first be labeled as change. Maps in near real-time systems evolve with each new satellite observation, and thus, the accuracies can fluctuate with time. We aim to assess all alerts generated within an evaluation period (e.g., 1 year), rather than the map at just a single instance of time.

The sampling design follows consolidated guidelines for assessing change maps using stratified random sampling (see Olofsson et al., 2014, and Chapter 3 of the current document). As the target change class in near real-time maps are often small relative to other map classes, a stratum to identify areas of omission is suggested; e.g., a spatial buffer around the alert class (Olofsson et al., 2020). If the near real-time system not only maps new possible changes but also removes from the map prior potential changes that were previously detected (due, for example, to the length of time since the initial detection or to reduced confidence), the areas of these removed change detections are valuable to consider in stratification as these areas are likely at higher risk of omission or commission error.

The response design builds on existing guidelines to collect reference labels. In addition to identifying the presence or absence of change, it is important to identify the timing of change, if present, with as much precision as the map being evaluated (e.g., date, minute). The source data used to collect reference labels should be of equal or better spatial and temporal resolution than what was used to generate the alert map. While *in situ* data at the sampling frequency of a near real-time map does not exist for most applications, dense satellite time-series imagery (often multi-sensor) can provide a valuable source of data to derive reference change status and timing through visual interpretation. However, gaps in data used for interpretation (due to satellite revisit rates, cloud cover, etc.) make it infeasible to interpret the precise timing when the land change occurred. In order to reduce temporal bias in the reference data, the reference time of change can be set to halfway between the previous clear observation and the observation in which change was first identified (Reiche et al., 2018). In addition to satellite imagery, near-surface cameras can serve as a reference data source for assessing the accuracy of detection of prompt land cover changes. They provide critical data on land dynamics, which is particularly valuable for validating near real-time maps. For instance, the global phenology monitoring network, PhenoCam (https://phenocam.nau.edu/webcam/), offers an example of such validation potential.

The analysis design is a set of procedures to quantify metrics of timeliness in addition to spatial accuracy metrics of omission and commission error of the change class with their associated measures of uncertainty. If the area of the no-change class is much larger than the change class, overall accuracy provides little insight into system performance.

Timeliness should be reported as an aggregate measure of time lag, which includes both observation lag and algorithm lag. Observation lag is calculated by the difference in the reference time and the map time. Algorithm lag requires additional information about the near real-time system or particular alert, as this represents the additional time from image acquisition to when the alert is integrated into the near real-time map. For some systems this may be a standard interval (e.g., it is always two days from acquisition to public availability of the source data and the system always incorporates the data the same day), while for others there may be variability due, for example, to a slower cadence of processing or to requiring a time-series of images after the first possible detection.

Example aggregate metrics of time lag include the mean/median time lag together with other metrics of the distribution (e.g., standard deviation) and associated confidence interval (Reiche et al., 2018) or similarly for other percentiles (e.g., 90% of alerts have a time lag less than X). For systems that update initial alert detections based on subsequent observations, a more complex approach can be to plot spatial accuracy versus time lag (Bullock et al., 2022). While various approaches may be used and these guidelines are merely preliminary, it is important to report on timeliness and spatial accuracy as both have substantial impacts on the usability of a system for a given application.

7.3 Toward more standardized validation datasets and collections of reference data

The principles for creating standardized reference datasets have been discussed in the literature (Olofsson et al., 2012; Stehman et al., 2012). The focus is on developing reference datasets that can be used to assess the accuracy of multiple land cover maps. The challenges related to this objective are in adapting sampling design to meet the accuracy estimation objectives for multiple maps and developing response design that can accommodate different land cover legends and spatial resolutions across validated maps. Stratified random sampling was proposed as a sampling design for standardized reference datasets as it meets the following criteria "*(1) it satisfies definition of a probability sampling design; (2) it provides adequate sample sizes for rare land-cover classes; (3) it allows flexibility to change sample size in response to unpredictable funding or revised accuracy assessment objectives; (4) it focuses sample sites in the areas most difficult for land-cover mapping*" (Olofsson et al., 2012). Accuracy assessment based on the reference datasets employing the stratified sampling designs needs to account for unequal inclusion probabilities of sample units in different strata by using the appropriate estimators weighting sample units according to their inclusion probabilities (see Chapter 5 for references to specific estimators). Response design was proposed to be flexible, with a number of core land cover classes standardized among the interpreters, and several optional sub-classes to enable some degree of customization within the general

reference classification framework. It was stressed that response design protocols for the collection of standardized reference datasets "*must be operationally practical and consistently implemented given that a large number of interpreters dispersed across the globe will likely be involved*" in the creation of such datasets (Olofsson et al., 2012). Standardization and detailed reporting of response design metadata (e.g., see Table 4.1) are therefore critical to ensure the quality of global standardized reference datasets. Specifically, if such a dataset is created by a large number of interpreters from around the world, critical details of the evaluation and labeling protocol (Table 4.1) need to be provided for each sample unit, e.g., who labeled each unit, how was the label assigned, level of certainty, the date of observation, etc.

Some of the **benefits** of creating standardized reference datasets include:

- *Consistency and comparability:* standardized datasets ensure that the products are validated based on uniform criteria, facilitating comparison between different products and studies (see section 2.8 for more discussion on comparative map accuracy assessment). Such consistency might help identify relative strengths and weaknesses in various land cover maps and enable the users to select the most appropriate map for their specific application.

- *Improved reliability:* a standardized validation dataset enhances accuracy assessment reliability. It provides a common reference, reducing biases and errors that may arise from using diverse and incompatible validation sources (e.g., multiple non-standardized regional datasets to validate a global map).

- *Resource optimization and cost reduction:* developing and maintaining a standardized dataset avoids duplication of effort and resources. Creating reference datasets is costly, especially when they rely on field visits to collect ground reference data, or on purchasing commercial very high-resolution imagery. Standardized datasets provide a common framework that can be used by multiple organizations, leading to more efficient use of time, effort and funding.

- *Monitoring, reporting and updates:* standardized datasets are crucial for monitoring changes in land cover and land use over time. They can enable Stage 4 validation (validation updates of operationally updated land cover maps, see section 7.1) or help reduce its cost, especially if regular revisits of sample units are performed as a part of the standardized reference dataset maintenance.

- *Transparent policy and decision making:* accurate and validated land cover and land use change maps are critical for informed policy-making and land management decisions. Standardized datasets ensure that these decisions are based on reliable and objective data. This also builds trust among users, including

scientists, policymakers, and the public, ensuring wider acceptance and utilization of these maps.

- *Capacity building and utilizing regional expert knowledge:* if the standardized validation datasets are developed globally or for large geographic regions, such initiative can support capacity building by providing local researchers and practitioners with high-quality data, at the same time enabling them to contribute to global research and decision-making processes and utilizing local and regional expertise.

Standardized reference datasets created and maintained by countries or international entities (see examples of the EU reference datasets below) can be general-purpose, characterizing all land cover classes (e.g., LUCAS dataset), or focus on a specific land cover or land use class (e.g., agriculture). The former are important for the validation of multi-class land cover maps but can be utilized for the validation of single-class land cover maps as well if this class is characterized with sufficient thematic detail in the general-purpose dataset. The latter are more suited for validating the specific land cover class maps and can accommodate more thematic detail but can also be utilized for assessing the class-specific accuracy of the multi-class maps. Geographic scale of the assessment also affects the level of thematic detail and quality of reference labels achievable in a standardized reference dataset. It is impossible to create a global standardized reference dataset that will satisfy the requirements of every application at every geographic scale. At the same time, global reference datasets can inform local and regional validation and reference data collection efforts, and local and regional reference datasets can support global map validation if the local reference data collection efforts are standardized enough (i.e., employ probability sampling designs, consistent land cover class definitions and data quality assurance procedures).

Standardization of reference datasets can also mean providing clear guidelines for reference data collection and metadata reporting, enabling regional entities and individual institutions to collect reference datasets that are potentially easier to integrate for validating maps covering larger areas. While the current protocol does not aim to establish a detailed protocol for creating standardized reference datasets, the core principles of sampling and response design discussed in Chapters 3 and 4 will apply to such datasets.

Some of the potential **challenges** that are likely to arise when creating standardized reference datasets are:

- *Quality of reference labels*: variability in the quality of reference data due to the differences in expertise and technology used in data collection. This can be addressed by implementing thorough quality control protocols and providing

training and resources to ensure data collectors adhere to high standards. Some of the approaches to ensuring the quality of reference data are discussed in Chapter 4 and in section 6.5.

- *Reference data availability*: unequal access to high-quality validation data, e.g., lack of global and temporally consistent very high-resolution satellite imagery coverage or difficulties obtaining field data in remote or politically unstable regions. Establishing international collaborations and funding mechanisms to support data collection in underrepresented areas and to make higher resolution datasets openly accessible (e.g., NICFI Planet mosaics funded by the government of Norway) is crucial to support reference data standardization efforts. Investing into acquiring spaceborne, airborne and UAV lidar data (see sections 6.2 - 6.3), airborne and UAV optical data and high-resolution radar data might help address inconsistencies in reference data availability in persistently cloudy areas.

- *Temporal discrepancies:* creation of standardized datasets is a long process, and differences in the timing of data collection can affect validation results, especially when relying on field campaigns or satellite data with inconsistent temporal coverage. Using protocols to harmonize data collection to a common timeframe and provide data collection (or reference imagery) dates in the metadata can help mitigate the effect of temporal inconsistencies.

- *Data formats and integration:* integrating data from multiple sources with different spatial resolution, formats and projections also can be challenging. It can be addressed by adopting interoperable data formats and conversion tools, resampling techniques and universal projections that all contributors must follow.

- *Funding, resources and organizational ownership:* the funding and resources to create, maintain and update standardized validation datasets are currently insufficient. The 3-year cycle common for funding research projects is not a viable model for supporting standardized reference datasets. International organizations, governmental agencies, and public-private partnerships dedicated to environmental monitoring should establish dedicated long-term funding with permanent staff and infrastructure dedicated to creation and maintenance of standardized reference datasets for land cover map validation, and to ensuring that these datasets are updated and provided to the public free of charge. The question of which institution oversees the process of creation of such a dataset and organizes inter-institutional collaboration is also related to funding and existing structure of national and international organizations and may be challenging to navigate.

- *Dissemination and adoption*: once the standardized reference datasets are created, users should be engaged through workshops, training sessions, and user feedback mechanisms to encourage widespread adoption and use.

- *Legal and ethical issues related to data privacy and ownership*: contributors' and authors' rights can limit data sharing. This can be solved by establishing clear data sharing agreements, anonymizing sensitive information, complying with international data privacy regulations and clear contributors' rights. Such agreements can be developed based on the experience of iNaturalist and other existing crowdsourced products. Regional institutions collecting field data need to be compensated for their work and appropriately acknowledged when publishing global findings based on compilations of regional datasets.

- *Preventing the use of standardized reference datasets for map training:* clear data use regulations and restrictions would be needed to prevent the use of the standardized reference datasets for map training, as this would compromise their independence from the maps being validated.

As a flagship example of standardized reference data collection, the European Union (EU) has developed key reference datasets: LUCAS, farmers' declarations, and Copernicus4GEOGLAM, each contributing uniquely to this domain.

LUCAS (Land Use/Cover Area frame Survey)

LUCAS, an EU initiative, collects harmonized land use and cover data across the EU. It employs systematic sampling (which is a probability sampling design, see section 3.3), providing a comprehensive overview of EU land cover types (Gallego and Delincé, 2010). This survey is instrumental in standardizing land cover data, enabling consistent land use assessments across EU member states. d'Andrimont et al. (2020) present a comprehensive effort to standardize and consolidate a vast array of *in situ* observations collected across the EU. Over five LUCAS surveys from 2006 to 2018, 1,351,293 observations were made at 651,780 unique locations, covering 106 variables and accompanied by 5.4 million photographs. The harmonization process described in the paper addresses the challenge of disparate datasets, unifying them into a singular, expansive database. This database offers an invaluable resource for geospatial and statistical analysis of land changes, with enhanced utility due to computational advancements like deep learning, providing an essential reference dataset for Earth Observation and contributing to more accurate characterization of land surface dynamics. Update of this dataset with the LUCAS 2022 dataset provides 400,000 new points, thus topping 1.7 million sample locations collected. The sampling methodology is described by Ballin et al. (2022) and the results of the survey can be obtained following the link (https://ec.europa.eu/eurostat/web/lucas/database/2022).

Since 2018, the LUCAS Copernicus module (d'Andrimont et al., 2021b) has provided advanced module support for Earth observation applications. It has seen a considerable increase in data collection, with 150,000 polygons gathered in 2022 (d'Andrimont et al., 2024), up from 60,000 in 2018, providing an extensive dataset for Earth observation applications. The 2022 data include 82 land cover and 40 land use classes, reflecting improvements in survey protocols and data collection efficiency.

Harmonized farmers declarations

The Geospatial Aid Application encompasses the crop declarations submitted by EU farmers for Common Agricultural Policy support, characterized by diverse methodological approaches across the EU member states. Current systems pose challenges in interoperability and semantic harmony, with limited public data access. Initiatives like AI4boundaries (d'Andrimont et al., 2023) and EuroCrops (Schneider et al., 2023) are pioneering the harmonization of parcel geometries and crop legends, with EuroCrops' open-source framework leading the community-driven effort toward semantic standardization. An EU regulation promises to revolutionize data availability, enhancing research and applications by providing public access to these high-value datasets.

Copernicus4GEOGLAM

Copernicus4GEOGLAM, a Copernicus service in support of the Group on Earth Observations Global Agricultural Monitoring Initiative (GEOGLAM), leverages the Copernicus program satellite data to support global agricultural monitoring. It provides validated crop monitoring baseline products, including crop type maps and crop area statistics, essential for food security early warning and response planning.

The main objective of the service is to strengthen national and sub-national level agricultural monitoring systems in GEOGLAM partner countries, by making available crop monitoring baseline products based on Sentinel 1 and 2 data. The service operated from 2020 on areas of interest (AOIs) between 100,000 and 200,000 km^2 in Uganda, Kenya, Tanzania and Ivory Coast. Extensive field campaigns are carried out in the AOIs and field survey data and mapping products are made publicly available at https://data.jrc.ec.europa.eu/collection/id-00356.The Copernicus4GEOGLAM service is aimed at maximizing the usability of ground-based datasets collected during various field campaigns by providing public access to georeferenced field observations and photographs, detailed text reports on methodologies and data. The service is currently funded up to 2028 and in 2024 it will next cover Yemen and Cameroon.

The EU's development of key reference datasets like LUCAS, harmonized farmers' declarations, and Copernicus4GEOGLAM underscore its commitment in standardizing Earth Observation data. This commitment enhances the accuracy and utility of datasets

for environmental and agricultural analyses, driving advancements in policy-making and scientific research.

Fiducial Reference Measurements (FRM)

An FRM is “a suite of independent, fully characterized, and traceable [to the International system of Units] [. . .] measurements of a satellite relevant measurand, tailored specifically to address the calibration/validation needs of a class of satellite borne sensor and that follow the guidelines outlined by the GEO/CEOS Quality Assurance framework for Earth Observation (QA4EO)” (Goryl et al., 2023). The concept of FRM provides evidence of the reliability of a reference dataset as a result of meeting a number of criteria. FRMs are attracting increasing interest from the satellite sensors’ and land product validation community and are mentioned for stage 4 validation in the CEOS validation hierarchy table (Figure 1.2.1).

Although the general principles and some mandatory criteria of FRM (e.g., independence from the satellite retrieval process, documented protocols, accessibility) overlap with those discussed above in relation to reference data for land cover map accuracy assessment, certain critical FRM criteria (e.g., traceability and comprehensive per-pixel uncertainty budget) were developed specifically for validating measurable quantities and cannot be easily transferred to categorical variables like land cover. Scientific efforts are ongoing to determine the extent to which FRM criteria can be applied to categorical variables derived from machine learning classification algorithms (Bilson et al., 2025). However, scientific consensus has not yet been reached, and collaboration between the metrological and land cover communities will be essential to establish an agreed approach to FRM for land cover. These developments will be documented in future updates of this land cover protocol.

7.4 Local map quality metrics

Chapters 3-5 focused on accuracy assessment and area estimation from probability samples of reference data, operating within design-based inference (see definition in section 2.1.1). In section 5.2, where model-assisted estimators of area were introduced, the statistical inference remains design-based. Design-based approaches avoid introducing subjectivity of the researcher and do not depend on a model. However, the metrics obtained by such an approach are global or aggregate estimates in the sense that they pertain to the entire sampled region of interest or a particular subregion as a whole.

Map users often require information about local map quality (Meyer and Pebesma, 2022). Such information can be provided, for example, by reporting the predicted probability of the label of a map unit to correspond to an independently observed

reference label at that same location. Such metrics are obtained by predictive models, and they are hence referred to as **model-based** predictions as opposed to the earlier described design-based estimates. Note that model-based methods do not require a probability sample to derive statistically valid inferences, although existing probability samples of reference data can be used in model-based methods, and in many instances may be preferable to non-probability samples (McRoberts et al., 2022).

Two common and contrasting approaches for obtaining local map quality metrics are:

1) Using internal metrics produced by the classification procedure itself, i.e., mapping uncertainty metrics such as the uncertainty in class assignment. This approach characterizes only disagreement between different model runs or models in reproducing the training data and therefore does not characterize the accuracy of the map.

2) Interpolation of agreement between the map and the reference labels obtained from an independent reference dataset. This approach is sometimes referred to as 'local accuracy assessment', because it aims to predict the probability of correct classification of map units, although it should be used only to supplement, and not to replace the recommended design-based estimates of overall- and class-specific accuracy metrics that characterize the entire mapped region or sub-region (section 5.1).

Map uncertainty metrics produced from classification outputs

The classification uncertainty approaches rely on the internal quality metrics of the classifier to assess the uncertainty of class assignment for each map unit. However, uncertainty metrics derived this way can only identify where the *classifiers* have most difficulty in distinguishing between the classes as represented by the training data, and not where the predicted class labels correspond to reality (Khatami et al., 2017; Stehman and Foody, 2019; Valle et al., 2023). Assessments with this approach have used maximum likelihood (Maselli et al., 1994), neural networks (Brown et al., 2009), random forests (Sales et al., 2022), ensemble methods (Witjes et al., 2022), resampling-based methods (Lyons et al., 2018) and conformal statistics (Valle et al., 2023). Foody (2022) showed that information from the classifier can be used to construct a complete confusion matrix. Alternatively, intercomparison of multiple land cover maps has been used for assessing local uncertainty in labeling (Gao et al., 2020). See section 2.9 for more examples of map intercomparison.

Interpolation approaches relying on sample data

As mentioned above, the interpolation approaches rely on an independent set of reference labels for a well-distributed sample that does not need not be collected via

probability sampling (Brus and De Gruijter, 1997). The agreement between the map and the reference labels known at the sample locations is spatially interpolated to provide predictions of the probability of correct labeling for each map unit (pixel) (e.g., Steele et al., 1998). Examples include Park et al. (2016), who integrated indicator-coded reference data with a wall-to-wall measure of classification uncertainty obtained from the classifier (see above) for geostatistical mapping of the probability of correct labeling. Khatami et al. (2017) used spatial kernels rather than geostatistics, and additionally considered spectral features, to predict the probability of correct classification. Other work has demonstrated the assessment of geographically-weighted correspondence matrices, from which local class-specific confusion probabilities can be calculated (Comber et al., 2017). Tsutsumida and Comber (2015) extended the approach of using geographically-weighted logistic regression to estimate local map quality to include the temporal dimension of accuracy.

Appendix. Examples of national-, regional- and global-scale validation efforts

This Appendix is meant to provide examples of large-scale (national to global) validation efforts that have been performed in the last two decades. As such, it might be a useful reference for map producers and users alike, to illustrate how some of the good practice recommendations discussed in Chapters 2-5 are implemented in practice. At the same time, this Appendix itself should not be viewed as a good practice protocol, as some of the individual studies might not follow all the recommended good practices.

A.1. Validation of the 'Global Land Cover 100m' from the Copernicus Global Land Service

The Copernicus Global Land Service – Dynamic Land Cover 100m (CGLS-LC100) product is a suite of land cover maps at a global scale with land cover characterized in discrete legends and fractions (Buchhorn et al., 2020a). This product was initially released in 2015 and was expanded to provide yearly land cover information from 2016 to 2019. Accordingly, independent reference data were collected to validate these maps for 2015 and were later expanded to allow yearly validation.

Copernicus Global Land Service Validation Data

The GCLS-LC100 validation dataset is based on a stratified sample using stratification by Olofsson et al. (2012). Here, each climate zone is divided into unpopulated and populated parts (more than 5 persons/km^2). Sample allocation per stratum and per continent was done by considering likely misclassification and landscape heterogeneity. Hence, more sample sites were allocated in heterogeneous areas such as the Sahel and dry savannah in Africa (Tsendbazar et al., 2018). To increase the sample representation in rare land cover types such as wetlands and water, additional sample sites were selected based on the GLS-LC100 V2.0 discrete land cover map with a minimum sample size requirement of 100 per land cover type for each continent (Tsendbazar et al., 2021). As a result, the global stratification consisted of 149 strata in total, divided over seven (sub)continents each with 19-25 strata, with 21,752 randomly selected sample sites (Figure A.1.4).

The reference land cover was visually interpreted using a dedicated web interface based on the Geo-Wiki platform (Figure A.1.1) (Fritz et al., 2012). Reference labels were interpreted by 30 regional experts, followed by revision and quality-checking processes. A sample unit covers an area of 100 m × 100 m that is divided into 100 small blocks/subpixels (10 m × 10 m) (Figure A.1.2) (Tsendbazar et al., 2018). Each subpixel is aligned to an individual Sentinel-2 L1C pixel at 10m (Buchhorn et al., 2020a). The

dominant land cover elements such as trees with different leaf and phenology types, shrubs, grass, crops, built-up areas, water, snow/ice, lichen/moss and regularly flooded areas were visually interpreted. The global validation data collection applied several steps to ensure good quality land cover reference data for validation (Figure A.1.3). After remote training, feedback was given in loops of interpretations (Tarko et al., 2021). Experts continued to the next loop when they had resolved the feedback on the previous loop received from validation experts (from Wageningen University). Feedback was given for each sample location. Next, consolidation steps were also done. Here, the reference land cover labels of the validation dataset were compared with national or regional land cover products such as Northern American Land Cover product (Latifovic et al., 2004), CORINE (Bossard et al., 2000), Australian Dynamic Land Cover, and Circumpolar Arctic Vegetation Map (Walker et al., 2005). Validation sites which did not match with these datasets were visually rechecked by the validation experts. For more details, see Tsendbazar et al. (2021).

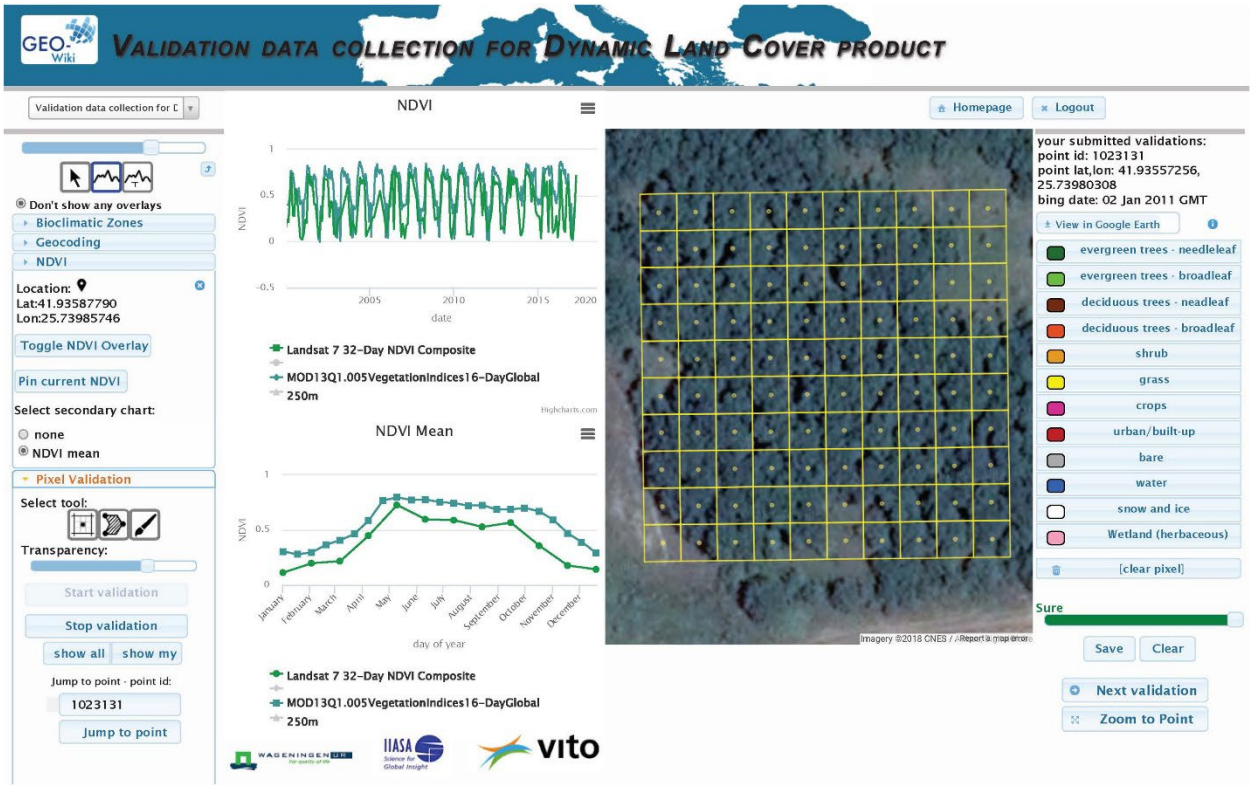

Figure A.1.1 Geo-Wiki based interface for land cover validation (figure from Tsendbazar et al., 2018). Overview of the Geo-Wiki platform is presented in Fritz et al. (2012).

Figure A.1.2 Example of sample interpretation over the 100 10mx10m sub-pixels included in a PROBA-V pixel (100mx100m). Figure from Tsendbazar et al. (2018).

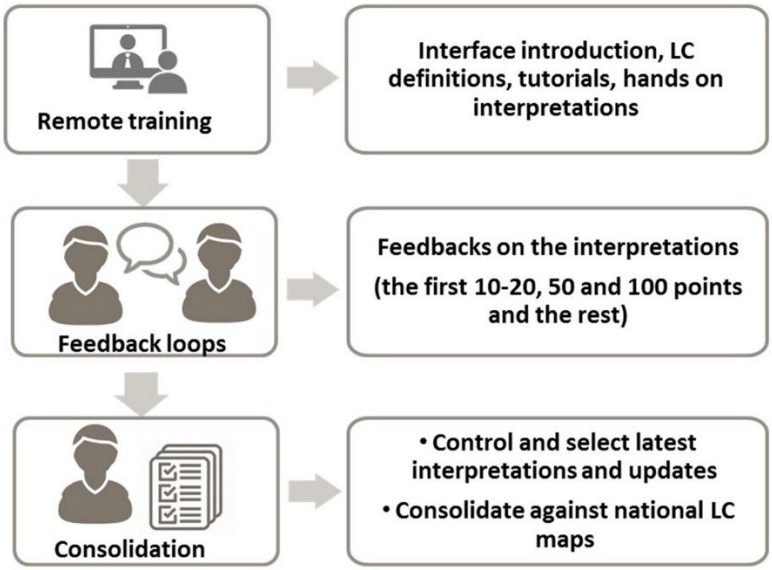

Figure A.1.3 Copernicus Global Land Service – Dynamic Land Cover 100m validation data collection and feedback process. Figure from Tsendbazar et al. (2021).

Validation data contained land cover labels at 10 m resolution subpixels and land cover fraction information of generic land cover elements at 100 m resolution at the sample site locations. At the 100 m level, the fraction information was translated into the CGLS-LC100 discrete map legend. Stratified estimators by Stehman (2014) were used to estimate map accuracy (the same estimators recommended in section 5.1). Furthermore, the land cover fraction maps were also assessed using these validation data to estimate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) per land cover type. For more details see Tsendbazar et al. (2021).

Updating the validation data

A first global land cover validation dataset was collected for the year 2015, and was then updated to the subsequent years, namely 2016 to 2019. Since the updates for these years were done at the same time, it is considered as one update. This was done based on the operational validation framework (see section 7.1).

First, a subset of the CGLS-LC100m validation dataset was rechecked on a randomly or targeted basis. Randomly selected 40% of the total sample sites were revisited for each continent over the update period (2016-2019), accounting for a 10% random revisit for each year. Next, sample sites with a high possibility of land cover change occurrence were also rechecked. Using the BFAST-Lite (Masiliunas et al., 2021) change detection algorithm on MODIS NIRv time series data, sample sites with 'breaks' were identified (6% of the total sample sites).

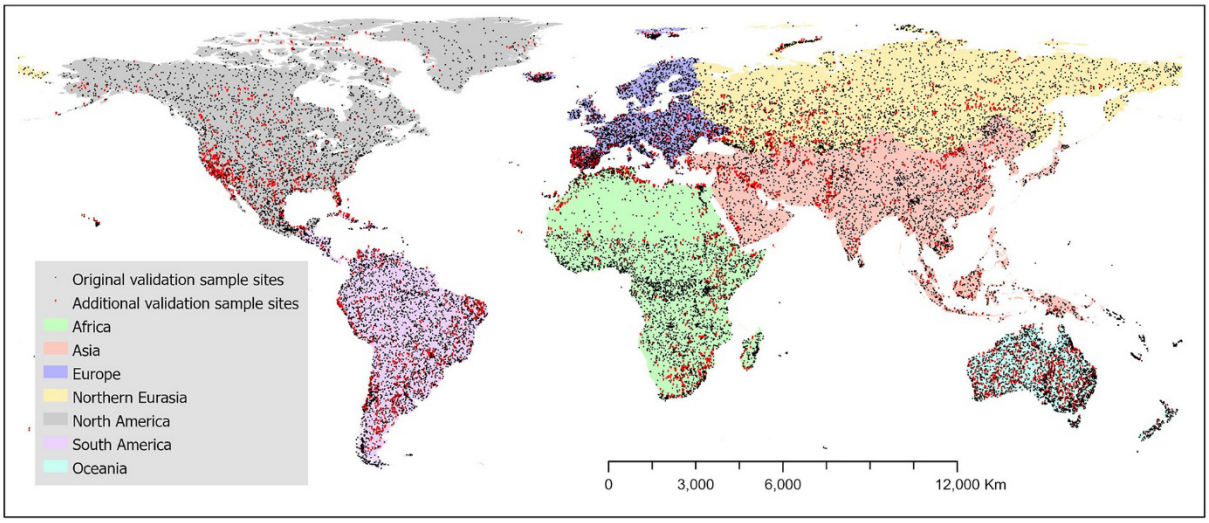

Figure A.1.4 (Sub) continental distribution of the validation sample sites: original (black squares) and additional (red squares). Figure from Tsendbazar et al. (2021).

Second, additional sample sites were selected and interpreted to better represent possible land cover change areas (Figure A.1.4). The possible land cover change stratification was created for each pair of years (2015-2016, 2016-2017, 2017-2018, and 2018-2019) using the annual CGLS-LC100 V3.0 land cover maps (Buchhorn et al., 2020b). To reduce spurious changes due to possible classification errors, this stratification was further adjusted using the break detection maps at a global scale and were limited to land cover change transitions that are deemed probable within the period and to areas that are at least 3 ha in size. Next, for each pair of years (e.g., 2015-2016), and for each continent, 240 sample sites were randomly selected based on the additional stratification. A total of 6720 sample sites for the update period (2015-2019) were collected. The original stratification was modified by imprinting the change stratification.

The areas of all strata and the sample inclusion probabilities were recalculated. For more details, see Tsendbazar et al. (2021).

The additional validation sample sites were visually interpreted by experienced experts involved in the global validation data collection. In the case of sample sites where no land cover change occurred, the reference land cover was labeled once. In the case of sample sites where land cover for at least one subpixel was changed, the land cover was labeled for each reference year in the period 2015-2019 accordingly (five times).To facilitate the interpretation of land cover and possible changes, where possible, multiple DigitalGlobe very high-resolution images per year were provided in addition to the Sentinel-2 time series thumbnails and time series NDVI profiles that are available in the Geo-Wiki platform (Figure A.1.1).

Following the accuracy assessment protocol of the validation dataset, the accuracy of the CGLS-LC100 V3.0 maps for 2015-2019 was then assessed based on the stratified estimators by Stehman (2014). For assessing the 2015 global land cover (GLC) discrete map, the original global validation data for 2015 was used. For the subsequent GLC maps (2016-2019), the updated global validation data which combines the original validation dataset with the additional validation dataset were used.

Lessons learned on operational validation of global land cover (GLC) maps

Although significant progress has been made in continuous land cover mapping at global scale, the continuous updating of validation data and validation estimates has been limited. Therefore, to inform map users about the quality of new releases and updates on land cover mapping, the importance of continuous validation as an essential part of all operational land cover mapping efforts is emphasized. Based on a framework for operational validation of GLC monitoring (see section 7.1), the CGLS-LC validation dataset and its update were used to validate the yearly CGLS maps from 2015-2019. This demonstrates GLC map validation in an operational context, and therefore one that meets the requirement of the stage 4 validation by the CEOS-LPV (see section 1.2 for the explanation of validation stages), focusing on updating validation for new releases or updates. Additionally, based on the multi-purpose nature of the validation dataset that can support the validation of maps at 10-100 m resolution, it was further updated to the years 2020 and 2021 to validate the ESA-WorldCover product at 10 m resolution (Zanaga et al., 2022). Validation designs with flexible spatial support and a sampling scheme that allows sample augmentation can be further focused to support efficient uses of validation datasets.

For updated map validation, keeping the validation dataset up to date is critical. This was addressed by partially rechecking the validation dataset, supplemented by augmenting sample sites in possible change areas within the operational validation of the

CGLS-LC100 maps. Considering the effort and time required for validation data collection, appropriate resources and funding support are required for any operational land cover monitoring efforts to update the validation dataset and improve sampling representation in land cover change areas.

Multi-temporal (annual and sub-annual) maps are often affected by the variability in classification results due to classifier uncertainty, and noise in the input data, which could lead to erroneous detection of land cover change (Sulla-Menashe et al., 2019). Temporal consistency of multi-year and continuous land cover products and the stability in their accuracy are therefore important for users interested in long term and consistent land cover observations (Bontemps et al., 2012). To address this, an assessment of the degree of stability in map accuracy has been conducted (Tsendbazar et al., 2021). Additionally, for users interested in land cover change information assessing the accuracy of land cover change is also important. When updating validation with new product releases, assessment of the accuracy, stability, and land cover change accuracy estimation are related yet somewhat different concepts that need to be considered for continuous and long-term land cover monitoring.

Finally, large-scale land cover monitoring is already progressing toward near-real-time and sub-annual frequencies, pioneered by the introduction of the Dynamic World product (Brown et al., 2022). This product brings out new aspects related to quality and fitness-for-purpose assessment. For example, to provide accuracy information with minimal delay, increasing the temporal precision in validation and reference data collection are the next steps to be developed.

A.2. ESA Climate Change Initiative global land cover time series

The ESA Climate Change Initiative C3S global land cover time series

The European Space Agency Climate Change Initiative (ESA CCI) Medium Resolution Land Cover (MR LC) project developed the methodology to produce state-of-the-art global land cover products that are consistent over long periods to contribute to the monitoring of Land Cover Essential Climate Variables. To achieve this goal, a series of global maps describing land cover and land cover change from 1992 to 2020 at a resolution of 300 m was generated using complete multi-mission Earth observation archives at both 300 m and 1 km (AVHRR, SPOT-Vegetation, MERIS, PROBA-V, and Sentinel-3) (Defourny et al., 2017). These maps are currently being produced within the framework of the Copernicus Climate Change service.

Accounting for the time dimension has enabled us to distinguish between the stable and dynamic components of land cover. What we refer to as land cover is defined as the baseline set of land cover characteristics that remain stable over time, independent

of any sources of seasonal, temporary, or natural variability (e.g., phenology) (Defourny et al., 2012). In contrast, land cover change is defined as a permanent change of the nature of land cover (e.g., deforestation). Conceptually and methodologically decoupling the land cover classification and land cover change detection ensures temporal and spatial consistency between successive maps.

The land surface is characterized by 22 classes using the Food and Agriculture Organization (FAO) Land Cover Classification System (Di Gregorio, 2005). Each year, thirteen types of land cover transitions are addressed, encompassing activities such as cropland and urban expansion, cropland abandonment, deforestation, water bodies drying up, etc. Detailed legends for both land cover and land cover change can be found in Defourny et al. (2017).

Validation methodology

The MRLC validation sampling scheme is based on the systematic sampling of the Joint Research Centre (JRC) TREES dataset, which is designed on a latitude/longitude geographical grid. This sampling approach is combined with a two-stage stratified cluster sample. The first stratum allows keeping an equal probability sampling based on the latitude. The second stratum reduces the sampling frequency among homogeneous landscapes (e.g., deserts) less prone to classification errors (Mayaux et al., 2006). A total of 2600 Primary Sampling Units (PSUs) were chosen from the entire sampling population (Figure A.2 (a)). This quantity represents a compromise between the time and effort needed for interpretation by land cover validation experts and the acceptable precision of the sample derived from a binomial distribution.

Figure A.2 (a) Spatial distribution of the 2600 Primary Sampling Units (PSU) from the CCI MRLC validation scheme; (b) Positioning of the Secondary Units within a PSU. Figure from Achard et al. (2011).

Five Secondary Sampling Units (SSUs) were systematically placed within a 20 km × 20 km box outlining each PSU. One SSU is situated at the centroid of the PSU, while

the remaining four are distributed at intervals of 4 km × 4 km from it. This arrangement of sample units, spaced at both the corners and center of a square, aims to minimize the spatial autocorrelation associated with the size of similar landscape elements. For the same reason, at each level, only 2 or 3 sample locations were selected among the 5 available. The SSU level served for the validation of the 300 m LC maps (Figure A.2 (b)).

An international network of land cover specialists with regional expertise and a comprehensive understanding of the CCI MRLC legend was responsible for building the MRLC validation database. For each SSU, visual interpretation of very high-resolution imagery close to the year 2010 was conducted to label pre-segmented objects. The evaluation of land cover change between 2010, 2005, and 2000 was systematically carried out on Landsat TM or ETM+ scenes acquired from the Global Land Surveys (GLS) for the respective years. Additionally, annual Normalized Difference Vegetation Index (NDVI) profiles calculated from SPOT-Vegetation time series aided in image interpretation by providing seasonal variations in vegetation greenness. Each SSU interpretation was assigned a level of confidence by the expert, categorized as either certain, reasonable, or doubtful. The homogeneity of each SSU, quantified as the area proportion occupied by the dominant class, is also documented.

This reference validation database was updated annually from 2016 to 2020 as part of the EU Copernicus Climate Change project, incorporating all available reference data at the time of validation for the years 2016, 2017, 2018, 2019, and 2020. Initially, validation sample units (SSUs) where the land cover was significantly impacted by major changes since 2010 (changes of at least 9 hectares) were identified. Subsequently, the land cover in these units was reinterpreted by experts.

Various accuracy parameters were generated and analyzed by comparing the land cover maps with independent interpretations by experts (ECMWF, 2020a). Confusion matrices were constructed, and overall accuracies were calculated, considering not only the diagonal cells representing correct classifications but also other cells indicating agreement between the product and the validation database. User and producer accuracies are also provided.

To avoid bias, one should ensure the independence of training and validation data (Strahler et al., 2006). Notably, it is widely accepted that accuracy assessments should avoid using the same sample data used for classification training. Radoux and Bogaert (2020) demonstrated that cross-validation introduces a bias in favor of the overall accuracy of the map under test. The systematic misrepresentations of ground conditions in the training set are propagated in the resulting land cover map. These errors remain unnoticed when a cross-validation set is subtracted from an erroneous training set. The procedure used here is validation in the strict sense, ensuring complete independence between the two types of data. This independence is facilitated by the quality of the

legend's description and its thorough understanding by both the photo interpreters and map producers.

Validation results

The evaluation was based on the dominant land cover class, homogeneity (>90%), and the expert confidence. Using ‘certain’ and ‘homogeneous’ SSU at 90%, overall accuracy values for the 2016 to 2020 land cover maps ranged from 70.5% to 71.1%. Notably, user accuracy was highest for cropland, broadleaf evergreen forest, urban, water bodies, and bare area classes. Disagreements were observed in classes like shrubland, grassland, and sparse vegetation (ECMWF, 2020b). Between 2016 and 2020, the land cover products showed varying rates of change in SSUs: 4.6% from 2010 to 2016, 0.3% from 2016 to 2017, 1.1% from 2017 to 2018, 1.2% from 2018 to 2019, and 0.85% from 2019 to 2020 (ECMWF, 2020b). These results indicate that land cover change is marginally distributed on a global scale. To increase sampling density in areas where change is more likely to occur, stratified random sampling should be designed to vary in space and time, targeting each land cover transition of interest (Olofsson et al., 2014). Such a sampling design and the associated validation database are currently under construction, and the map accuracies obtained with this new design will be published as soon as possible.

A.3. ESA Climate Change Initiative Water Bodies product

Multiple cartographic products often map the same class, making it essential to assess the relative quality of a product that is, in relation to other products, in addition to its intrinsic quality. Although there are clear guidelines explaining how to assess the accuracy of a given product (this document), little experience is reported in product comparison frameworks. The overall accuracy of a map sometimes fails to distinguish between similar products, particularly in the case of strongly imbalanced binary maps (e.g., water bodies, deforestation). In the context of the ESA CCI MR LC project (already introduced in section A.2), our goal was to tackle challenges associated with the joint comparison and validation of binary map products on a global scale, specifically focusing on scenarios where one class, such as water bodies, is marginally distributed (Lamarche et al., 2017). We selected three global inland water products: the CCI WB product v4.0 (Lamarche et al., 2017), the SAR-WBI (Santoro and Wegmüller, 2014), the GIW v1.0 (Feng et al., 2016a) dataset and the GFC-datamask (Hansen et al., 2013) and performed their accuracy assessment using a reference database created from 2110 photo-interpreted sample sites of very high-resolution imagery.

We used a stratified random sampling approach with strata corresponding to error-prone areas in water body mapping. We compared this stratification with other sampling methods such as simple random sampling or class-based stratified random sampling

(strata defined using a map of the target class). The confidence-based stratification was divided into three strata (Figure A.3): high confidence in correctly mapping the land class (Stratum 1), high confidence in correctly mapping the water class (Stratum 2), and error-prone areas (Stratum 3). The surface area of Stratum 3, i.e., error-prone areas, accounted for 76% of the total inland water surface.

The sample size was determined using Equation 3.1 in section 3.5. We operated under the assumption that the accuracy of water classification is lower in areas where different maps show disagreement, whereas water bodies are typically classified with high to very high overall accuracy in areas of agreement. Consequently, the desired overall accuracy in the error-prone area was set to be at least 85%, requiring approximately 1200 sample units with a width of the confidence interval set to 4% with a confidence level of 95% (z = 1.96). To evaluate overall accuracy within the two strata characterized by higher product agreement and consequently higher expected accuracy, an additional 1,200 sample units were evenly distributed among these two strata. Thus, we distributed half of the total of 2400 sample units in error-prone areas and allocated one-fourth (600 units) each to the land agreement and water agreement strata with the aim to have the smallest SE in areas where the products disagree. Although somewhat arbitrary, this allocation was found realistic given time and resource constraints. Using Equation 3.1, the expected overall accuracy therefore corresponds to 93%.

The sampling unit was the pixel materialized with a footprint of 150 m x 150 m. These sample pixels were visually interpreted independently of the product, using high-resolution Google Earth imagery. Particular care was taken to interpret and record the permanent and temporary characteristics of snow and water uniformly across the globe, using numerous historical images. According to the photo-interpretation practices building on the convergence of evidence (Estes and Simonett, 1975), it was possible to identify the presence of water at the time of imaging, as well as surfaces that may be seasonally flooded. Specifically, these surfaces include dry riverbeds, flood-prone areas, irrigated agriculture, mangroves/inundated forests, ephemeral streams, salt pans, and snow packs. Sample pixels were labeled as water when at least half of the sample was covered with open surface water. Sample pixels displaying temporary snow or water were labeled as land, with the temporal aspect also being recorded. For all sample pixels, the date of the high-resolution imagery was documented. Additionally, wetlands and swamps were recorded.

Of the original 2400 sample pixels, 279 pixels were discarded because they corresponded to invalid data in at least one of the spatially incomplete datasets used to build CCI WB v4.0. Eleven sample pixels were subsequently excluded, either due to challenging interpretation of Google Earth imagery caused by cloud coverage, unavailability of images, or uncertain interpretation. Of the 2110 remaining sample pixels, 1030 were incorporated into the error-prone stratum, and 234 corresponded to temporary

water bodies such as ephemeral streams, beaches, irrigated crops, and salty lakes. Figure A.3 shows the reference database comprising 2110 distributed across land masses, excluding polar areas.

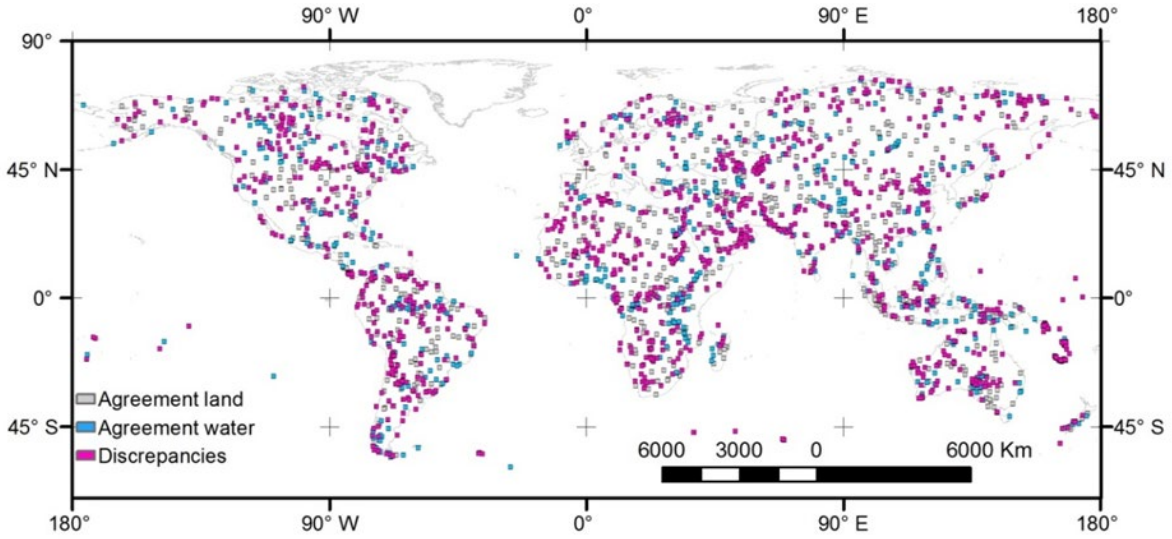

Figure A.3 Location of the 2110 sample sites (area of 150 m x 150 m) selected for the validation of the water body products in 3 strata defined following an approach targeting error-prone areas. Figure from Lamarche et al. (2017).

The quality of the three global inland water maps products was assessed on the basis of confusion matrices constructed by comparing each map class with the classes in the reference sample. These matrices were used to calculate overall accuracy (OA), user accuracy (UA), producer accuracy (PA) (Strahler et al., 2006) and the F1 score (section 5.1). The overall accuracy of all three products (the SAR-WBI, the GIW v1.0 and the GFC-datamask) was always very high, between 98 and 100 %. This was a consequence of the overwhelming proportion of the land class at the global scale compared to the marginal water class. However, water surfaces were not consistently identified with high accuracy in any of the input datasets. The PA for water exhibited significant variations among the individual input datasets. The PA of the CCI WB v4.0 (92%) surpassed the highest PA among the input datasets by 13%. Conversely, the UA of the GFC-datamask was higher (97%) compared to the UA of the CCI WB v4.0 (86%). The GFC-datamask exhibited water omissions primarily along lake banks, shallow water, and dams. The minimum water extent in the GFC-datamask is likely attributed to its definition of water, using a strict threshold of equal to or greater than 50% of water detections in the Landsat image time series (Potapov et al., 2012). Commission errors are marginal and occur along some lakes and over black lava rocks (e.g., Ethiopia).

During the estimation of class area, sampling methods are typically compared by examining their standard errors or the confidence interval of the estimated area. Using

equations 10 and 11 from Olofsson et al. (2014) we have estimated a SE of the inland water area of 17.5 mln. km² for simple random sampling, 12.5 mln. km² for class-based sampling, and 12.9 mln. km² for confidence-based sampling. Simple random sampling clearly results in the highest SE (lowest precision), corroborating its inefficiency to assess surfaces of marginally distributed classes. However, while oversampling the error-prone area is beneficial for comparing classifiers it is not ideal for precise area estimation, as evidenced by the larger SE in confidence-based sampling compared to class-based sampling.

In the context of area estimation, it would be beneficial for precision of the estimates to have larger sample size in Stratum 1, since it covers over 95% of the total area. A practical compromise could involve allocating half of the sample units to stratum 1 and the rest - split equally to Strata 2 and 3. This would lead to a 22% smaller relative SE of area estimates but a larger 40% relative SE for the product intercomparison. This new sample design would be better than the class-based design to both estimate areas (10 mln. km² SE vs 12.9 mln. km²) and compare products.

A.4. UMD GLAD validation of single- and multi-class land cover and change maps

University of Maryland's (UMD's) Global Land Analysis and Discovery (GLAD) laboratory (https://glad.umd.edu/) has over a decade of experience in assessing the accuracy of land cover and change maps and estimating the area of land cover and change classes from a reference sample. Most of UMD GLAD's projects are national to global scale, which poses additional challenges, outlined below.

Sampling design considerations: stratification

Typically, UMD GLAD's validation activities employ stratified sampling with strata constructed from maps derived from medium-resolution satellite imagery (MODIS and Landsat). Broich et al. (2009) demonstrated the utility of MODIS-based (500m resolution) stratification for estimating the area of deforestation in the Brazilian Legal Amazon from 30m Landsat imagery, and Pickering et al. (2019) the benefits of Landsat-based stratification when using higher resolution reference data (5m RapidEye) for estimating the area of forest loss in the high forested low deforestation country of Guyana.

Most of the UMD GLAD's studies use Landsat-resolution stratification and Landsat as a primary source of reference data, which is consistent with existing good practice guidelines (Olofsson et al., 2014). Typically, the map that is being validated is used for stratification to target land cover class of interest, but the strata do not have to match the map classes. Maps of relatively rare classes, such as land cover change, are generally conservative, meaning that the target class is somewhat underestimated to reduce the

amount of noise in the final map. Considering this, a buffer stratum is often added around such mapped classes to target potential errors of omission frequently occurring on the class boundaries (Hansen et al., 2016; Pickens et al., 2020; Potapov et al., 2022a; Turubanova et al., 2023, 2018; Tyukavina et al., 2022, 2018, 2017, 2015, 2013; Ying et al., 2017). A 'possible target class omission' stratum could be defined some other way, e.g., based on the existing independent maps or classification algorithm uncertainty metrics (Potapov et al., 2022b). Adding such strata helps better quantify map omission errors and mitigates the variance inflation effect that map omission errors from large strata have on the area estimates (Olofsson et al., 2020). Map commission errors could affect area and accuracy estimates of larger classes, for which maps used for stratification are less conservative (e.g., forest extent). This effect could be mitigated by splitting the large target stratum into the 'core', where confusion between mapped and reference classes is less likely, and the 'periphery' or 'buffer inside' (e.g., inside the forest from the forest edge) from the boundary of the mapped class (Potapov et al., 2017, 2015; Turubanova et al., 2023).

In addition to sampling strata related to the land cover class being mapped or estimated, regions for which individual accuracy metrics or area estimates are needed to be reported are often sampled separately, to ensure that each reporting region has adequate sampling density. Examples of this are sampling regions based on biomes and climate domains (Hansen et al., 2013; Ying et al., 2017), continents and subcontinents (Potapov et al., 2022b; Tyukavina et al., 2022), countries (Turubanova et al., 2018), carbon density strata (Tyukavina et al., 2015). In other cases, reporting regions are defined as subpopulations or subdomains similar to post-strata (Tyukavina et al., 2018, 2017). If the sample size within a reporting region is too small to yield the desired precision of the estimates, it may be necessary to augment the sample.

Stratification for validating a single class land cover (e.g., forest vs. no forest) or land cover change map (e.g., forest loss vs. no forest loss) is relatively straightforward, and usually includes the target class stratum or strata, the non-target class stratum, and one or more 'probable target class' strata. Stratification for validating multi-class land cover maps or estimating the area of multiple land cover classes from the sample is more challenging. For static multi-class land cover map validation (Hansen et al., 2022) and sample-based estimation of the area of multiple classes (Potapov et al., 2017; Zalles et al., 2021), the validation can be based on a single stratification that includes separate strata for the most important classes or the classes most likely to have high errors (e.g., wetlands or change classes) while minimizing the overall number of strata (e.g., combining into one stratum land cover change classes small in area with the static map of the same class). Another option when estimating map accuracy and area of multiple land cover and change classes combined in a single map is the simple random sampling (Potapov et al., 2019), although such an approach can result in higher standard errors of

the accuracy metrics for the small land cover and change classes. An alternative approach for validating land cover and change maps with a large number of classes is using a separate set of strata and a separate sample to validate each land cover change theme, e.g., the dynamics of forest extent, cropland, water, built-up, and perennial ice and snow (Potapov et al., 2022a). This approach focuses on user's and producer's accuracies of individual land cover and change themes mapped separately (and later combined into a multi-class land cover and change map) instead of reporting the accuracy of the combined map.

Sampling design considerations: sampling unit

Most UMD GLAD's studies employ a 30x30m Landsat pixel as a primary sampling unit, to match the primary mapping unit of the maps that are being validated. Point sampling is used when working with reference data stored in different grids, e.g., 30m Lat/Long (geographic coordinates) grid of GLAD Analysis Ready Data (ARD) Landsat data and 10m UTM grid of Sentinel-2 (Pickens et al., 2022). When only Landsat data is available for all sample units, each 30x30m pixel is assigned with binary reference labels (yes/no target class). When higher resolution data is available for most sample units, sub-pixel class proportions can be estimated (Potapov et al., 2017). Using larger blocks of pixels as sampling units, e.g., 5x5, 12x12, 20x20 or 24x24 km equal-area pixel blocks (Khan et al., 2016; Pickering et al., 2021, 2019; Potapov et al., 2014; Song et al., 2021) allows optimizing the acquisition of high-resolution satellite or ground reference data and employing model-assisted estimators (e.g., difference or regression) with wall-to-wall auxiliary information from existing maps to reduce standard errors of the area estimates (Stehman, 2013, 2009b). Larger blocks of pixels are more challenging to interpret in the context of multi-class assessments, hence they are primarily used for the single-class validation and area estimation. Blocks of pixels can be mapped to derive proportions of target land cover classes (Pickering et al., 2021, 2019) resulting in one-stage cluster sampling, or a sample of points or pixels can be interpreted or visited in the field for each sample block (Khan et al., 2016; Potapov et al., 2014; Song et al., 2021) resulting in a two-stage cluster sampling.

Sampling pixels in geographic coordinates

The global GLAD ARD Landsat-based dataset (Potapov et al., 2020) and all resulting maps are stored in a 0.00025°degree pixel grid in geographic coordinates. In such a grid, the area on the ground that each pixel represents gets smaller moving from the equator to the poles. The difference in pixel size is often negligible if working at the national scale, particularly for the countries located close to the equator, and thus sampling of GLAD ARD pixels can be treated as sampling of equal-area units.

Conversely, in the global studies, the difference between the equatorial and the polar pixel size is significant, which would result in over-representation of the polar areas in the sample. Various strategies have been developed to adjust for this effect, namely sampling pixels with inclusion probabilities proportional to their area (Pickens et al., 2022, 2022, 2020; Tyukavina et al., 2022), creating climate domain sampling strata with small within-stratum variability of pixel size (Potapov et al., 2022b; Ying et al., 2017), or reprojecting the map being validated into an equal-area projection (Hansen et al., 2013). Summarizing these approaches, Tyukavina et al. (2025) present a unified set of estimators that allows performing equal probability sampling of GLAD ARD pixels and accounting for varying pixel area at the estimation stage, or sampling pixels with inclusion probabilities proportional to their area.

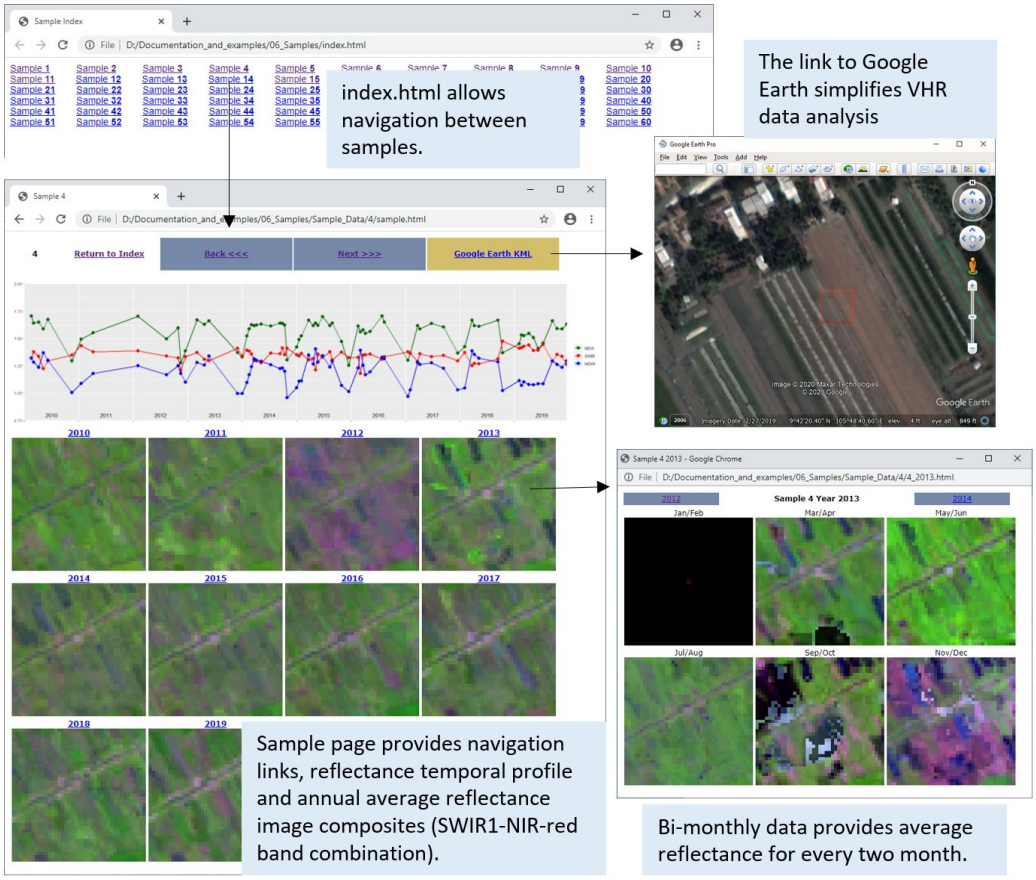

Figure A.4 Example of the reference data web page for a sample pixel (red box) from the GLAD ARD User Manual, available at https://glad.umd.edu/ard/home

Sampling and reference data visualization tools

GLAD ARD toolset (https://glad.umd.edu/ard/home) provides a manual and necessary scripts for simple random or stratified sampling of ARD pixels for the purposes of map accuracy assessment and area estimation. The toolkit includes the scripts for creating sampling strata from mapping results (including creating buffers around mapped

classes) and rasterized vectors of reporting regions (countries, biomes, etc.). It also provides tools for allocating the sample to the sampling strata. Once the reference sample is selected, a specific tool allows to build HTML pages with sample reference data (Figure A.4), including annual Landsat mosaics and 16-day observations from GLAD ARD data, plots of spectral index time-series based on the 16-day data, and a link to Google Earth for each sample pixel. GLAD ARD tools also include scripts for estimating map accuracy and land cover class area, based on the equations that are applicable regardless of whether the sampling strata match map classes (Stehman, 2014).

A.5. Validation of the European crop map 2018

The EU crop map (d'Andrimont et al., 2021a) is the first attempt to consistently map cropland at the parcel level over the entire European Union (EU). It is generated by combining meaningful information from the Land Use/Cover Area frame Survey (LUCAS) survey with Synthetic Aperture Radar (SAR) Sentinel-1 (S1) data for the year 2018. Specifically, the study capitalizes on the unique LUCAS 2018 Copernicus module, introduced to collect *in situ* data fitting Earth Observation (EO) processing requirements information on a specific subset of LUCAS 2018 points. The final 10-m map for 2018, detecting 19 different crop types is available for download and visualization at: https://data.jrc.ec. europa.eu/dataset/15f86c84-eae1-4723-8e00-c1b35c8f56b9.

In d'Andrimont et al. (2021a), three approaches and datasets are tested to estimate the accuracy of the EU crop map. The first approach is taking as reference data the high-quality LUCAS core points not surveyed by the Copernicus module. The second approach is comparing the EU crop map with a selection of Geospatial Aid Application (GSAA) data based on farmers' declarations. The third approach compares the area of several main crops obtained from the EU crop map to the corresponding official subnational statistics. Here we will focus on the use of the LUCAS data 2018 for the accuracy assessment.

The 2018 LUCAS survey

In the EU, *in situ* data collection is organized through the LUCAS (https://ec.europa.eu/eurostat/web/lucas). For the description of the LUCAS survey please refer to section 7.3.

The 2018 LUCAS core points

The survey consists of a two-phase sampling. In the first phase, 1.1 million geo-referenced points are systematically drawn forming a 2x2 km^2 grid, i.e., one point every 2 km in the EU. The points are then stratified according to land cover classes to allow the second phase of sampling. In 2018, this resulted in 337,854 LUCAS core points for which

97 variables were collected by surveyors in the field or by photo interpretation in the office. Please note that the LUCAS core variables are the ones collected for each point surveyed (i.e., the identification of the point, and the surveying of specific variables on different aspects of land cover, land use, and land and water management; Eurostat, 2018). In addition to the core variables, some specific modules could be collected providing additional specific information.

The 2018 LUCAS Copernicus Module

Despite LUCAS being designed for EU-wide standardized reporting of land cover and land use area statistics and not for EO applications (see section 7.3), LUCAS *in situ* data are increasingly used in land cover and land use EO research. In addition, a new LUCAS module specifically tailored to EO was introduced in 2018: the LUCAS Copernicus module. A specific protocol was designed to collect *in situ* information with specific characteristics fitting EO processing requirements. Specifically, the LUCAS Copernicus module collected the exact geolocation of the observation as well as information on the spatial extent and homogeneous continuity of the land cover observed around the surveyed LUCAS point (Figure A.5.1). It helps overcoming the limitations related to the difference of scale between a decametric pixel size and a 1.5 m circle radius around the LUCAS core point as well as inaccuracies in geo-location of the observation.

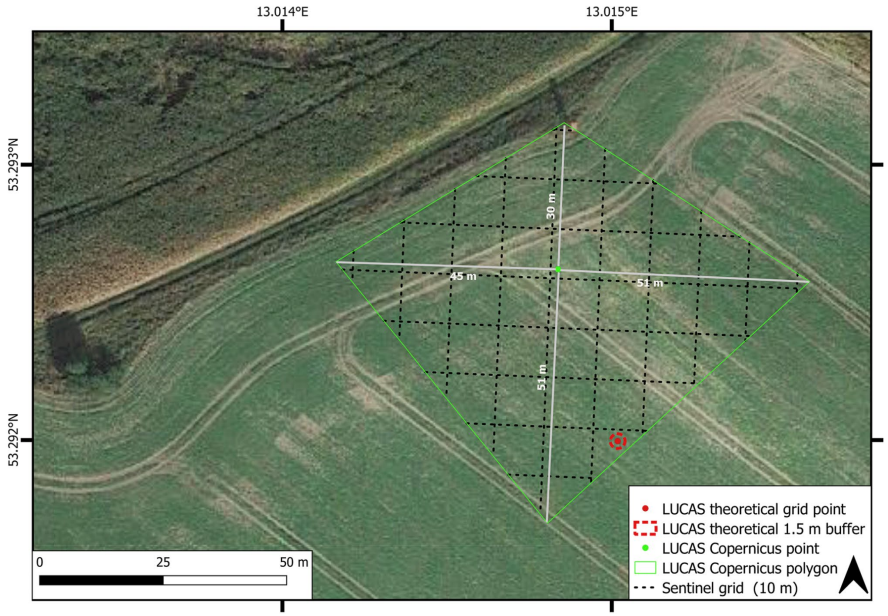

Figure A.5.1 Building the Copernicus polygon geometry. Figure from d'Andrimont et al. (2021b).

The Copernicus module was planned for 90,620 points and actually executed for 63,364 points. The sample design of the module could be viewed as a two-phase sample design, with the LUCAS sample representing the first phase and the second phase allocated according to land cover classes probabilities. In addition, only LUCAS core

points visited *in situ* are eligible for inclusion in the Copernicus module (Ballin et al., 2022, 2018). Post-processing the Copernicus data collected with the steps presented in d'Andrimont et al. (2021b), a total of 58,428 polygons with homogeneous land cover (extending up to 0.52 ha) are retrieved with a level-3 land cover (66 specific classes including crop type) and land use (38 classes) information.

Validation of the EU crop map 2018 with LUCAS core points

Due to the limited amount of Copernicus polygons, all the polygons have been allocated to the training of the classifiers used to create the EU crop map 2018. However, the accuracy assessment could be done at the EU scale level with the LUCAS core points filtered to keep only high-quality information. Four criteria are applied to keep only direct and *in situ* observations, remove parcels smaller than 0.1 ha and only keep data with homogeneous land cover. In addition, the subset of (63,364) points surveyed for the COPERNICUS module are filtered out. This screening resulted in a total of 87,853 LUCAS core points (spatial distribution overview in Figure A.5.2).

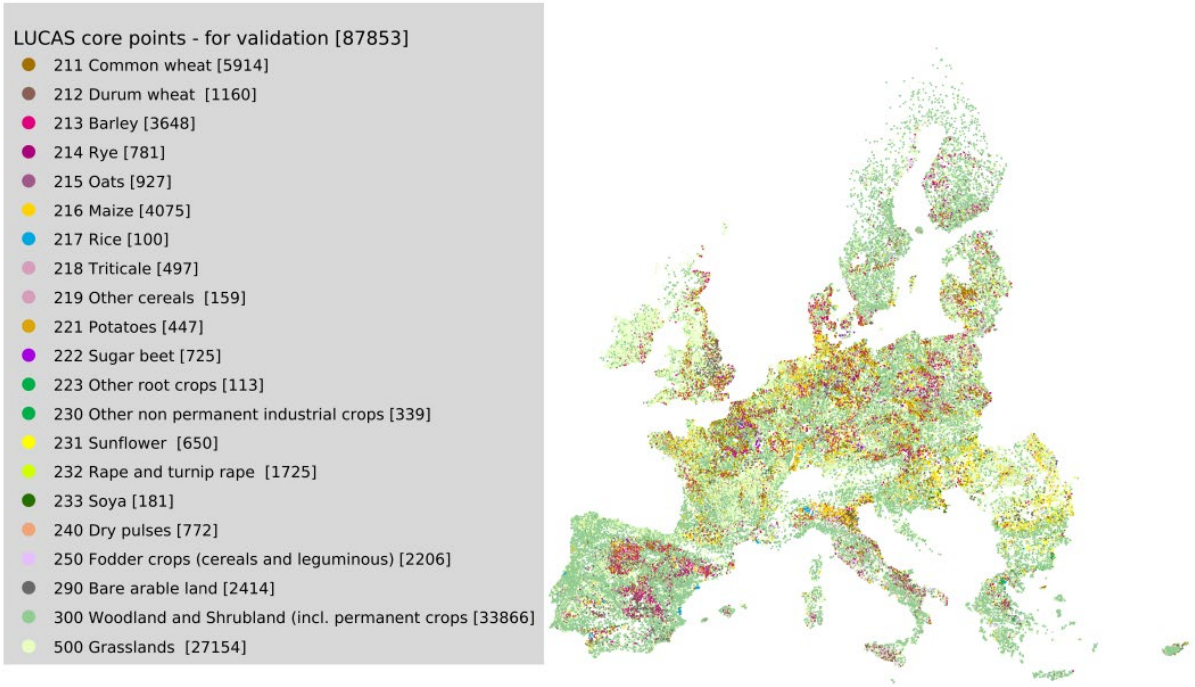

Figure A.5.2 Validation points (87,853) are high-quality points of the LUCAS core survey 2018 not used for training. Figure from d'Andrimont et al. (2021a).

The reference data are used in combination with the EU crop map to report the confusion matrix, the overall, user, and producer accuracy. LUCAS is a two-phase sampling scheme. The LUCAS 2018 survey followed a stratified random sample design (Ballin et al., 2018). The first phase is a systematic sampling scheme, and we treat the second phase, the 2018 sample sites, as collected under a stratified random sampling,

due to the large amount of collected data at the first phase. Accordingly, the accuracy metrics are reported based on the estimated area proportion of correctly and mis-classified classes according to the equations of Olofsson et al. (2014) for a 95% confidence level. The accuracy for the main crops is reported in Table A.5.

A novel approach is tested in Verhegghen et al. (2021) to derive a local accuracy assessment. The 2018 high-quality LUCAS core points are used to evaluate the 'per-pixel land cover accuracy' (PLCA) of the EU crop map, following Ebrahimy et al. (2021). With this approach, the LUCAS reference dataset is converted to a binary validation dataset, in which each LUCAS point is assigned with 1 if correctly classified in the EU crop map and 0 if incorrectly classified. Random forest (RF) classifiers are then used to establish a nonlinear relationship between the binary reference dataset and the S1 time series used for the classification. A RF is built for each class of the EU crop map and the PCLA is predicted using the probabilistic output of the RF model. Besides providing local accuracy information, this type of approach is interesting when the validation sample is not a probability sample and therefore does not allow design-based inference.

Table A.5 Overall accuracy for the crop type classification and producer's (PA) and user's accuracies (UA) for the main crop type classes, with standard errors. Accuracies and standard errors are expressed as proportions (0 to 1); to convert to percentages multiply by 100. Table from d'Andrimont et al. (2021a).

Class	PA	SE	UA	SE
common wheat	0.78	0.01	0.50	0.01
durum wheat	0.26	0.02	0.50	0.03
barley	0.54	0.01	0.52	0.02
maize	0.84	0.01	0.59	0.01
potatoes	0.37	0.03	0.74	0.06
sugar beet	0.57	0.03	0.75	0.03
sunflower	0.87	0.02	0.62	0.03
rape and turnip rape	0.83	0.02	0.80	0.02
Overall accuracy	0.76	0.003		

Although the original LUCAS survey is not planned for validation of EO derived products, this exercise shows that the dataset can be used in meaningful ways to provide an accuracy assessment of a land use map over Europe.

The EU Crop Map 2022 and the LUCAS 2022 In Situ Survey

Building on the foundational work of the 2018 EU crop mapping effort, the 2022 update presents a refined and comprehensive view of Europe's agricultural landscapes. Leveraging the latest advancements in Earth Observation technology and methodologies, the EU Crop Map 2022 (European Commission, 2022) extends its predecessor's ambition by offering a more detailed and accurate portrayal of crop distribution across the

European Union and Ukraine. This iteration benefits significantly from the enriched dataset provided by the LUCAS 2022 *in situ* survey. The 2022 crop map employs a sophisticated methodology that integrates LUCAS Copernicus polygons, Sentinel satellite imagery, Land Surface Temperature data, and a Digital Elevation Model. This approach, mirroring the successful combination of LUCAS survey data and Sentinel-1 SAR data from 2018, takes advantage of the increased number of Copernicus module points and the simplified survey protocol of the LUCAS 2022 survey. The result is a high-resolution (10-meter) Land Use Land Cover (LULC) map with improved classification accuracy for 19 specific crop types. This map, developed through a Random Forest machine learning algorithm, showcases an overall accuracy of 79.3% for major land cover classes and 70.6% for the crop types, marking a significant step forward in the precision and utility of agricultural mapping in Europe.

Furthermore, the LUCAS Copernicus 2022 survey's (d'Andrimont et al., 2024) expanded dataset, including around 150,000 polygons, reflects a substantial increase from the 63,364 points collected in 2018. This growth not only enhances the training and validation of the crop map's classification models but also underscores the evolving capabilities and contributions of *in situ* surveys to EO-based agricultural monitoring. The EU Crop Map 2022 and the LUCAS 2022 survey together represent a pivotal advancement in our understanding of European agriculture, offering invaluable insights for sustainable farming practices, food security planning, and environmental policy making.

A.6. Validation activities within the Satellite Observatory of Central African Forests (OSFAC) context

The Satellite Observatory of Central African Forests is a non-governmental organization with a regional vocation which has been working in Central Africa for twenty years, monitoring the forests of the Congo Basin through satellite data (https://osfac.net/fr/). For the validation of land cover and land use products, OSFAC has supported several projects and programs of national institutions of Central African countries and international organizations such as the Central Africa Regional Program for the Environment of the United States Agency for International Development, UMD, NASA, World Resource Institute (WRI), European Union Forest Institute (EFI, https://efi.int/), World Bank, Food and Agriculture Organization of the United Nations (FAO), United Nations Environment Programme (UNEP), United Nations Development Programme (UNDP), etc. The methodological approaches used for map validation at OSFAC are diverse.

Figure A.6.1 Field data collection with Teleported Aerial System (RPAS) in the DRC. Figure by Jörg Haarpaintner, NORUT, Norway.

In the framework of an OSFAC and NORUT (Northern research Institute Norway) collaboration, forest maps produced by remote sensing were validated using an error matrix approach with reference data coming most often from very high-resolution satellite imagery, in accordance with the method and the GFOI Guidance Document (GFOI, 2020). FAO tools such as Open Foris Collect and Collect Earth were also used by OSFAC for map validation. In addition, high-resolution photographs taken from a teleported aerial system (RPAS) were used as a source of reference data (Figure A.6.1).

In another collaboration, OSFAC was responsible for the systematic quality review of national maps of Central African forests at 10 m resolution using Sentinel satellites produced by the Catholic University of Louvain (UCLouvain), particularly those of the Congo. All the polygons of the map were systematically reviewed according to a regular grid (see section 2.10 for more details on the systematic quality review), using all available auxiliary data. Verification of forest classes was based on the visual interpretation of the shape, texture, local environment, and color (colored composition) of the entities making up each land cover class in high-resolution imagery (Figure A.6.2).

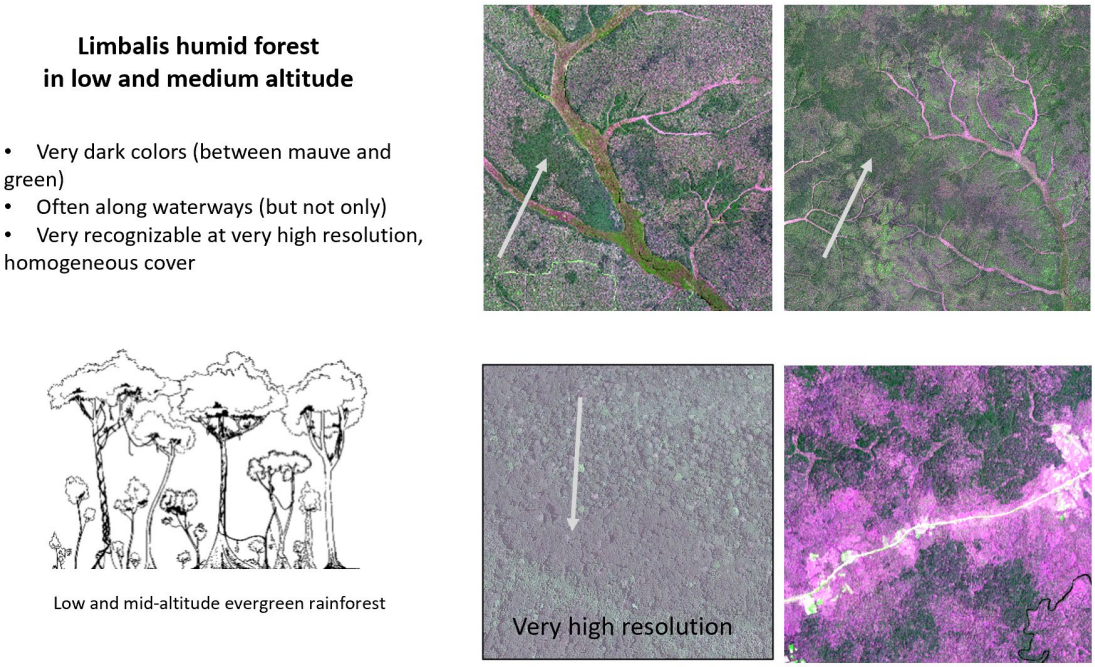

Figure A.6.2 Example of visual identification of forest cover type in satellite imagery by texture, color, and landscape context. Imagery is from Google Earth, data attribution: Google, Maxar, Airbus. Figure by Landing Mané.

As part of the Development of the Reference Emission Level (REL) for Unplanned Deforestation in the Maï-Ndombe (Democratic Republic of Congo) Emission Reductions Program Area (Wildlife Works) program, OSFAC validated the forest change mapped for over a decade. The validation methodology consisted first in assessing the accuracy of the detection of forest change classes: Unplanned Deforestation, Degradation and Afforestation/Reforestation. Sample sites were randomly selected within the boundary of the selected change strata (stratified random sampling). At least three experts were responsible for interpreting the sample sites. The sample sites which had the agreement of all the experts were accepted after the initial interpretation, while the other ones with disagreements were reinterpreted by groups of experts with the aim of finding a consensus. At the end, a confusion matrix was created, and the accuracy of the classification result calculated (see Chapter 5 and Table 5.1.1 for the general methodology description).

As a last example, OSFAC validated the map of the vegetation classes of the Miombo Forests in the province of Haut-Katanga in the Democratic Republic of Congo. The probability sampling for map accuracy assessment was performed with the aim to collect reference data in the field, which was supplemented with visual interpretation of very high-resolution imagery from Google Earth (Figure A.6.3).

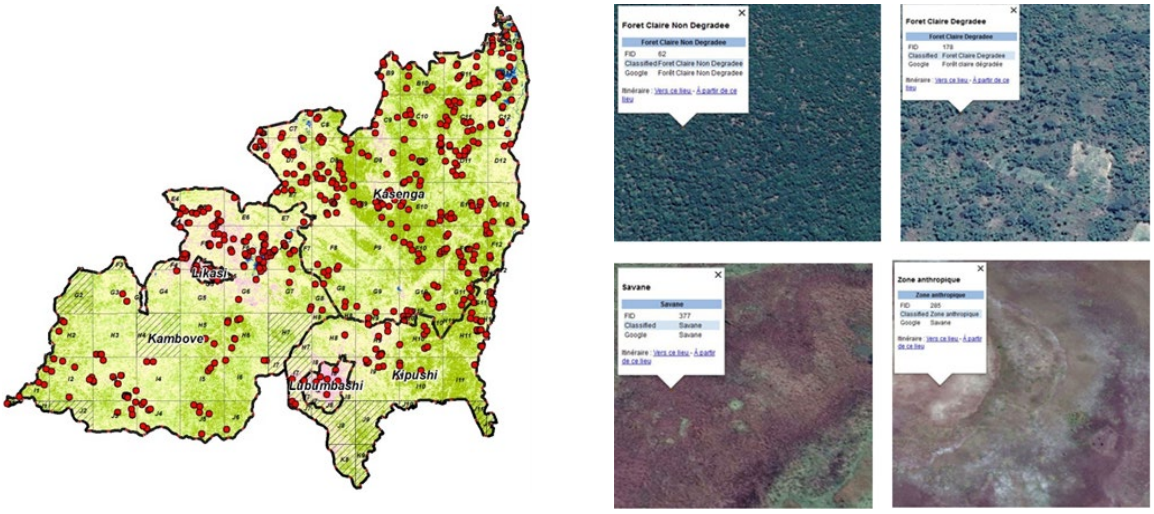

Figure A.6.3 Sample of points used to validate land cover classification in Katanga Province DRC (left) and the use of high-resolution images from Google Earth to assign reference land cover classes (right). Imagery attribution: Google, Maxar, Airbus. Figure by Landing Mané.

These examples of validation of land cover and land use products show that the approaches are diverse. At OSFAC, the opinions of map producers and users are taken into account in the choice of the validation method to be applied for each project or program.

A.7. Validation strategy for land cover and land cover change in support of GLanCE

The goal of the ***Gl***obal **Lan**d ***C***over ***E***stimation (GLanCE) project is to create global maps of land cover and land cover change for the period 2000-2020 at 30 m spatial resolution from Landsat (Friedl et al., 2022). For the GLanCE datasets to be useful, their accuracy and uncertainty need to be quantified based on high-quality reference data allocated via probability sampling (Olofsson et al., 2014; Stehman, 2000). To this end, the GLanCE project is compiling reference data to estimate: (1) the overall accuracy of land cover and land cover change datasets being created by the project; and (2) the user's and producer's accuracy, along with estimates of associated 95% confidence intervals, for each map category.

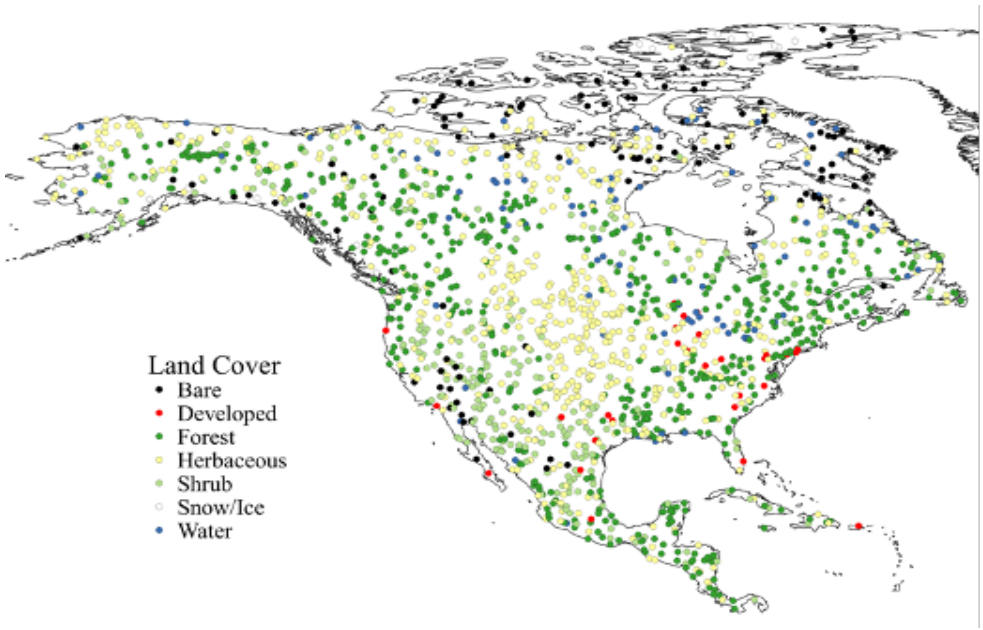

Figure A.7 Map showing locations of reference sites selected via random sampling in North America in support of land cover accuracy assessment for the GLanCE product. Figure from Friedl et al. (2022).

To support rigorous error and uncertainty estimation, two sets of random samples are being collected. The first sample is being collected to characterize the accuracy of GLanCE land cover data in each year of the record and is based on a probability sample drawn from all the Earth's land areas (Figure A.7). The second sample is being collected to support characterization of accuracy and uncertainty in land cover change. At global scale, land cover change impacts a small proportion of the land surface in any given year. Hence, this latter dataset is selected by stratified random based on stable versus changed land areas. In both cases, the design-based inference framework that GLanCE is using for accuracy assessment requires high quality reference observations. To this end, each reference site is independently interpreted by multiple analysts. Any sites for which consensus across interpreters cannot be achieved are removed from the sample and replaced with new sites using the original sampling design; less than 2% of sample units for each continent were removed and replaced in this way. To account for bias in area estimates from pixel counting, stratified estimators (described in Olofsson et al., 2013) are used for area estimation.

A.8. Validation of the cropland maps

Accurate monitoring of the distribution and types of crops and reliable forecasting of crop production can provide critical information to effectively manage national and global food supply. Recent advances in Earth Observation systems, data availability and cloud computing facilitate our ability to monitor crops (Becker-Reshef et al., 2020;

Whitcraft et al., 2019; Wu et al., 2023). With the growing interest in thematic products characterizing the distribution and changes in agricultural land cover, cropland extent and spatially explicit crop type, several groups around the world are producing such data products (Defourny et al., 2019; Fritz et al., 2019; Nakalembe et al., 2021; Potapov et al., 2022b). However, the utility of the products depends in large part on their accuracy. In that context, the Committee on Earth Observation Satellites (CEOS) Land Product Validation (LPV) Working Group on Land Cover (CEOS LPV WG LC) and the Group on Earth Observations Global Agricultural Monitoring (GEOGLAM) initiative joined forces to develop community good practices for accuracy assessment of cropland and crop type satellite-derived product accuracy. This section summarizes their recommendations at the time of writing these guidelines.

Most global land cover products include a cropland class, with sometimes an additional discrimination between annual and permanent crops and/or information about irrigation status, while spatially explicit crop type information mainly exists at sub-continental and national scale. Producing validated accurate crop products requires addressing a set of underestimated challenges specific to this land cover class:

1. Importance of dating cropland and crop type maps, along with the associated reference data and accuracy estimates to account for cropland and crop seasonality: cultivated area is changing over the year and between years. A diversity of crop calendars often co-exists over the same region, including the distinction between annual and permanent crops.

2. Intensive field campaigns are required to gather enough high-quality reference data to validate the maps, in addition to producing them. While simple cropland vs. non-cropland reference data can be obtained by visual interpretation of very high-resolution (VHR) imagery, this is not possible for differentiating complex non-cover classes (like grassland, pastures, fallow, etc.) and for crop type. Several field campaigns are possibly needed to address the variety of crop calendars.

3. High variability in space and over time for the same crop makes the collection even more challenging (e.g., crop density, pouring cereal, crop disease, grazing).

4. A variety of agricultural practices introduce an additional level of variability at the crop type level: organic farming, cover crop, tillage, fertilization, etc. Irrigation is a key practice for its impact on water use, which introduces specific crop calendars and growing cycles.

The validation approach will mainly depend on the objective of the data producer and/or on the expected use of the map. Nevertheless, a set of minimum requirements have been identified that should be used as key principles when establishing the validation protocol, which largely agree with the general principles of land cover map

accuracy assessment (see section 2.2). First and foremost, probability sampling is a must. It is possible to also use data collected through other opportunistic sampling, but they need to be kept separate unless their inclusion probabilities are known.

Second, the quality of the collected reference data needs to be assessed. Errors in ground data are large even if they are assumed to be error-free and these errors may lead to misinterpretation of accuracy. Using Artificial Intelligence models to provide labels for reference data is not recommended. Ideally, labels are to be provided by teams of experts and their level of certainty are to be recorded and taken into account. Ideally, reference data should be shared.

Third, there are a variety of response design approaches that can be useful and efficient. It is therefore important to document the approach through metadata. Reference data can be collected through field visits, visual interpretation of VHR imagery, UAV flights, pictures on the ground, etc. Spatial support regions can be points, polygons, blocks, or segments. Labels can represent the majority area of the spatial support or provide the crop classes proportions. The protocol can focus on the dominant label or also include secondary ones. The complexity becomes even larger when discussing the way to handle mixed cropping, stressed, damaged or failed crops, etc.

In terms of accuracy reporting, basic metrics are encouraged like user's and producer's accuracy while kappa index is clearly discouraged (see sections 2.7 and 5.1 of the current protocol). The confusion matrix should be provided in area proportions (i.e. adjusted for sample design). Cropland and crop type area can be estimated from the same reference sample used for accuracy assessment with little to no additional effort. An honest reporting is recommended to highlight the strength of the map products, but also to document deviations in the sampling and response design (e.g., sample site not visited) transparently.

As stated previously, these key principles can be tailored to specific objectives, and it is important to define these objectives before designing the validation protocol. Figure A.8 shows a high-level roadmap aiming to support the definition of such protocol, following the list of minimum requirements listed above. Finally, like for other domains, accuracy assessment needs to be taken seriously, and a substantial budget needs to be allocated to this task.

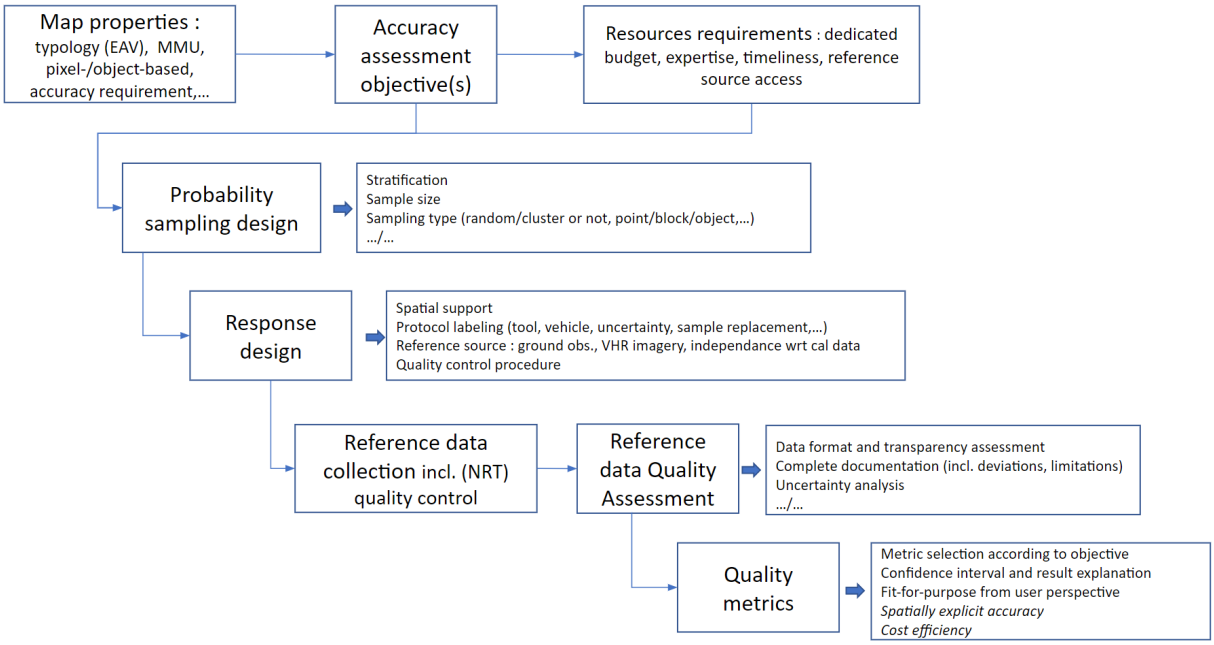

Figure A.8 Roadmap to establish an efficient scientifically valid protocol for validation crop maps. Figure by Pierre Defourny and Sophie Bontemps.

References

Achard, F., Bontemps, S., Defourny, P., Herold, M., Mayaux, P., 2011. Land Cover CCI Product validation plan.

Achard, F., Stibig, H., Eva, H., Mayaux, P., 2002. Tropical forest cover monitoring in the humid tropics-TREES project. Tropical Ecology 43, 9–20.

Ahlqvist, O., 2005. Using uncertain conceptual spaces to translate between land cover categories. International Journal of Geographical Information Science 19, 831–857. https://doi.org/10.1080/13658810500106729

Al-Najjar, H.A.H., Kalantar, B., Pradhan, B., Saeidi, V., Halin, A.A., Ueda, N., Mansor, S., 2019. Land Cover Classification from fused DSM and UAV Images Using Convolutional Neural Networks. Remote Sensing 2019, Vol. 11, Page 1461 11, 1461. https://doi.org/10.3390/RS11121461

Alvarez-Vanhard, E., Corpetti, T., Houet, T., 2021. UAV & satellite synergies for optical remote sensing applications: A literature review. Science of Remote Sensing 3, 100019. https://doi.org/10.1016/J.SRS.2021.100019

Aristeidou, M., Herodotou, C., Ballard, H.L., Young, A.N., Miller, A.E., Higgins, L., Johnson, R.F., 2021. Exploring the participation of young citizen scientists in scientific research: The case of iNaturalist. PLOS ONE 16, e0245682. https://doi.org/10.1371/JOURNAL.PONE.0245682

Asokan, A., Anitha, J., 2019. Change detection techniques for remote sensing applications: a survey. Earth Science Informatics 12, 143–160. https://doi.org/10.1007/S12145-019-00380-5/FIGURES/6

Ballin, M., Barcaroli, G., Masselli, M., 2022. New LUCAS 2022 Sample and Subsamples Design: Criticalities and Solutions. Eurostat statistical working papers, Publications Office of the European Union: Luxembourg. Publications Office of the European Union, Luxembourg. https://doi.org/10.2785/957524

Ballin, M., Barcaroli, G., Masselli, M., Scarnó, M., 2018. Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018. Eurostat statistical working papers. Publications Office of the European Union, Luxembourg.

Bassine, C., Radoux, J., Beaumont, B., Grippa, T., Lennert, M., Champagne, C., De Vroey, M., Martinet, A., Bouchez, O., Deffense, N., Hallot, E., Wolff, E., Defourny, P., 2020. First 1-M Resolution Land Cover Map Labeling the Overlap in the 3rd Dimension: The 2018 Map for Wallonia. Data 2020, Vol. 5, Page 117 5, 117. https://doi.org/10.3390/DATA5040117

Bastin, J., Berrahmouni, N., Grainger, A., Maniatis, D., Mollicone, D., Moore, R., Patriarca, C., Picard, N., Sparrow, B., Abraham, E.M., Aloui, K., Atesoglu, A., Attore, F., Bey, A., Garzuglia, M., García-montero, L.G., Groot, N., Guerin, G., Laestadius, L., Lowe, A.J., Mamane, B., 2017. The extent of forest in dryland biomes. Science 638, 1–5.

Bayas, J.C.L., See, L., Bartl, H., Sturn, T., Karner, M., Fraisl, D., Moorthy, I., Busch, M., van der Velde, M., Fritz, S., 2020. Crowdsourcing LUCAS: Citizens Generating Reference Land Cover and Land Use Data with a Mobile App. Land 2020, Vol. 9, Page 446 9, 446. https://doi.org/10.3390/LAND9110446

Becker-Reshef, I., Barker, B., Humber, M., Puricelli, E., Sanchez, A., Sahajpal, R., McGaughey, K., Justice, C., Baruth, B., Wu, B., Prakash, A., Abdolreza, A., Jarvis, I., 2019. The GEOGLAM crop monitor for AMIS: Assessing crop conditions in the context of global markets. Global Food Security 23, 173–181. https://doi.org/10.1016/J.GFS.2019.04.010

Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., Verdin, J., 2020.

Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. Remote Sensing of Environment 237, 111553. https://doi.org/10.1016/J.RSE.2019.111553
Berger, K., Machwitz, M., Kycko, M., Kefauver, S.C., Van Wittenberghe, S., Gerhards, M., Verrelst, J., Atzberger, C., van der Tol, C., Damm, A., Rascher, U., Herrmann, I., Paz, V.S., Fahrner, S., Pieruschka, R., Prikaziuk, E., Buchaillot, M.L., Halabuk, A., Celesti, M., Koren, G., Gormus, E.T., Rossini, M., Foerster, M., Siegmann, B., Abdelbaki, A., Tagliabue, G., Hank, T., Darvishzadeh, R., Aasen, H., Garcia, M., Pôças, I., Bandopadhyay, S., Sulis, M., Tomelleri, E., Rozenstein, O., Filchev, L., Stancile, G., Schlerf, M., 2022. Multi-sensor spectral synergies for crop stress detection and monitoring in the optical domain: A review. Remote Sensing of Environment 280, 113198. https://doi.org/10.1016/J.RSE.2022.113198
Bey, A., Díaz, A.S.P., Maniatis, D., Marchi, G., Mollicone, D., Ricci, S., Bastin, J.F., Moore, R., Federici, S., Rezende, M., Patriarca, C., Turia, R., Gamoga, G., Abe, H., Kaidong, E., Miceli, G., 2016. Collect earth: Land use and land cover assessment through augmented visual interpretation. Remote Sensing 8, 1–24. https://doi.org/10.3390/rs8100807
Bian, L., 2007. Object-Oriented Representation of Environmental Phenomena: Is Everything Best Represented as an Object? Annals of the Association of American Geographers 97, 267–281. https://doi.org/10.1111/J.1467-8306.2007.00535.X
Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., Bontemps, S., Leroy, M., Achard, F., Herold, M., Ranera, F., Arino, O., 2008. GlobCover Products Description and Validation Report.
Bilson, S., Cox, M., Pustogvar, A., Thompson, A., 2025. A metrological framework for uncertainty evaluation in machine learning classification models. https://doi.org/10.48550/arXiv.2504.03359
Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. A fuzzy set-based accuracy assessment of soft classification. Pattern Recognition Letters 20, 935–948. https://doi.org/10.1016/S0167-8655(99)00061-6
Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., Shirk, J., 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. BioScience 59, 977–984. https://doi.org/10.1525/BIO.2009.59.11.9
Bontemps, S., Defourny, P., Van Bogaert, E., Arino, O., Kalogirou, V., Ramos Perez, J., 2011. GlobCover 2009 Products Description and Validation Report.
Bontemps, S., Herold, M., Kooistra, L., Van Groenestijn, A., Hartley, A., Arino, O., Moreau, I., Defourny, P., 2012. Revisiting land cover observation to address the needs of the climate modeling community. Biogeosciences 9, 2145–2157. https://doi.org/10.5194/BG-9-2145-2012
Boogaard, H., Pratihast, A.K., Bayas, J.C.L., Karanam, S., Fritz, S., Van Tricht, K., Degerickxi, J., Gilliams, S., 2023. Building a community-based open harmonised reference data repository for global crop mapping. PLOS ONE 18, e0287731. https://doi.org/10.1371/JOURNAL.PONE.0287731
Bossard, M., Feranec, J., Oahel, J., 2000. CORINE land cover technical guide - Addendum 2000. European Environment Agency.
Bourgoin, C., Ameztoy, I., Verhegghen, A., Desclée, B., Carboni, S., Bastin, J.-F., Beuchle, R., Brink, A., Defourny, P., Delhez, B., Fritz, S., Gond, V., Herold, M., Lamarche, C., Mansuy, N., Mollicone, D., Oom, D., Peedell, S., San-Miguel, J., Colditz, R.R., Achard, F., 2024. Mapping global forest cover of the year 2020 to support the EU regulation on deforestation-free supply chains. Publications Office of the European Union.
Broich, M., Hansen, M.C., Potapov, P., Adusei, B., Lindquist, E., Stehman, S.V., 2011. Time-series analysis of multi-resolution optical imagery for quantifying forest cover loss in

Sumatra and Kalimantan, Indonesia. International Journal of Applied Earth Observation and Geoinformation 13, 277–291. https://doi.org/10.1016/j.jag.2010.11.004
Broich, M., Stehman, S.V., Hansen, M.C., Potapov, P., Shimabukuro, Y.E., 2009. A comparison of sampling designs for estimating deforestation from Landsat imagery: A case study of the Brazilian Legal Amazon. Remote Sensing of Environment 113, 2448–2454. https://doi.org/10.1016/j.rse.2009.07.011
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., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., Tait, A.M., 2022. Dynamic World, Near real-time global 10 m land use land cover mapping. Scientific Data 9. https://doi.org/10.1038/s41597-022-01307-4
Brown, K.M., Foody, G.M., Atkinson, P.M., 2009. Estimating per-pixel thematic uncertainty in remote sensing classifications. International Journal of Remote Sensing 30, 209–229. https://doi.org/10.1080/01431160802290568
Brus, D.J., De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44. https://doi.org/10.1016/S0016-7061(97)00072-4
Buchhorn, M., Lesiv, M., Tsendbazar, N.E., Herold, M., Bertels, L., Smets, B., 2020a. Copernicus global land cover layers-collection 2. Remote Sensing 12. https://doi.org/10.3390/rs12061044
Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.E., Herold, M., Fritz, S., 2020b. Copernicus Global Land Service: Land Cover 100m: collection 3: epochs 2015-2019: Globe [Dataset].
Bullock, E.L., Healey, S.P., Yang, Z., Houborg, R., Gorelick, N., Tang, X., Andrianirina, C., 2022. Timeliness in forest change monitoring: A new assessment framework demonstrated using Sentinel-1 and a continuous change detection algorithm. Remote Sensing of Environment 276, 113043. https://doi.org/10.1016/J.RSE.2022.113043
Burnicki, A.C., 2011. Modeling the probability of misclassification in a map of land cover change. Photogrammetric Engineering and Remote Sensing 77, 39–50. https://doi.org/10.14358/PERS.77.1.39
Burrough, P.A., 1996. Natural Objects with Indeterminate Boundaries, in: Geographic Objects with Indeterminate Boundaries. CRC Press, pp. 3–28. https://doi.org/10.1201/9781003062660-2
Bwangoy, J.R.B., Hansen, M.C., Roy, D.P., Grandi, G.D., Justice, C.O., 2010. Wetland mapping in the Congo Basin using optical and radar remotely sensed data and derived topographical indices. Remote Sensing of Environment 114, 73–86. https://doi.org/10.1016/j.rse.2009.08.004
Câmara, G., 2020. On the semantics of big Earth observation data for land classification. Journal of Spatial Information Science 21–34.
Card, D.H., 1982. Using Known Map Category Marginal Frequencies to Improve Estimates of Thematic Map Accuracy. Photogrammetric Engineering and Remote Sensing 48, 431–439.
Carlotto, M.J., 2009. Effect of errors in ground truth on classification accuracy. International Journal of Remote Sensing 30, 4831–4849. https://doi.org/10.1080/01431160802672864
Chazdon, R.L., Brancalion, P.H.S., Laestadius, L., Bennett-Curry, A., Buckingham, K., Kumar, C., Moll-Rocek, J., Vieira, I.C.G., Wilson, S.J., 2016. When is a forest a forest? Forest concepts and definitions in the era of forest and landscape restoration. Ambio 45, 538–550. https://doi.org/10.1007/s13280-016-0772-y
Chen, Jun, Chen, Jin, Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30 m resolution: A

POK-based operational approach. ISPRS Journal of Photogrammetry and Remote Sensing 103, 7–27. https://doi.org/10.1016/j.isprsjprs.2014.09.002
Christen, P., Hand, D.J., Kirielle, N., 2023. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. ACM Computing Surveys 56. https://doi.org/10.1145/3606367/ASSET/91B75E27-EDA5-45BA-AFA6-D20864C7414A/ASSETS/GRAPHIC/CSUR-2022-0380-F08.JPG
Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.C., Skakun, S.V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sensing of Environment 219, 145–161. https://doi.org/10.1016/j.rse.2018.09.002
Cochran, W.G., 1977. Sampling techniques, 3rd ed. John Wiley & Sons, Inc., New York.
Cohen, W.B., Yang, Z., Kennedy, R., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync - Tools for calibration and validation. Remote Sensing of Environment 114, 2911–2924. https://doi.org/10.1016/j.rse.2010.07.010
Colditz, R.R., Schmidt, M., Conrad, C., Hansen, M.C., Dech, S., 2011. Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions. Remote Sensing of Environment 115, 3264–3275. https://doi.org/10.1016/J.RSE.2011.07.010
Comber, A., Brunsdon, C., Charlton, M., Harris, P., 2017. Geographically weighted correspondence matrices for local error reporting and change analyses: mapping the spatial distribution of errors and change. Remote Sensing Letters 8, 234–243. https://doi.org/10.1080/2150704X.2016.1258126
Comber, A., Fisher, P., Brunsdon, C., Khmag, A., 2012. Spatial analysis of remote sensing image classification accuracy. Remote Sensing of Environment 127, 237–246. https://doi.org/10.1016/J.RSE.2012.09.005
Comber, A., Fisher, P., Wadsworth, R., 2005. What is Land Cover? Environment and Planning B: Urban Analytics and City Science 32, 199–209. https://doi.org/10.1068/B31135
Comber, A., Tsutsumida, N., 2023. Geographically weighted accuracy for hard and soft land cover classifications: 5 approaches with coded illustrations. International Journal of Remote Sensing 44, 6233–6257. https://doi.org/10.1080/01431161.2023.2264503
Comber, A.J., Wadsworth, R.A., Fisher, P.F., 2008. Using semantics to clarify the conceptual confusion between land cover and land use: the example of ‘forest.’ Journal of Land Use Science 3, 185–198. https://doi.org/10.1080/17474230802434187
Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35–46. https://doi.org/10.1016/0034-4257(91)90048-B
Congalton, R.G., Green, K., 2019. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Third Edition. Assessing the Accuracy of Remotely Sensed Data. https://doi.org/10.1201/9780429052729
Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., Lambin, E., 2004. Digital change detection methods in ecosystem monitoring: A review. International Journal of Remote Sensing 25, 1565–1596. https://doi.org/10.1080/0143116031000101675
Cushman, S.A., Gutzweiler, K., Evans, J.S., McGarigal, K., 2010. The Gradient Paradigm: A Conceptual and Analytical Framework for Landscape Ecology. Spatial Complexity, Informatics, and Wildlife Conservation 9784431877714, 83–108. https://doi.org/10.1007/978-4-431-87771-4_5
Czaplewski, R.L., 2003. Accuracy Assessment of Maps of Forest Condition: Statistical design and methodological considerations., in: Wulder, M., Franklin, S. (Eds.), Remote Sensing of Forest Environments. Springer, Boston, MA, pp. 115–140. https://doi.org/10.1007/978-1-4615-0306-4_5

d'Andrimont, R., Claverie, M., Kempeneers, P., Muraro, D., Yordanov, M., Peressutti, D., Batič, M., Waldner, F., 2023. AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography. Earth System Science Data 15, 317–329. https://doi.org/10.5194/ESSD-15-317-2023

d'Andrimont, R., Verhegghen, A., Lemoine, G., Kempeneers, P., Meroni, M., van der Velde, M., 2021a. From parcel to continental scale – A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations. Remote Sensing of Environment 266, 112708. https://doi.org/10.1016/J.RSE.2021.112708

d'Andrimont, R., Verhegghen, A., Meroni, M., Lemoine, G., Strobl, P., Eiselt, B., Yordanov, M., Martinez-Sanchez, L., Van Der Velde, M., 2021b. LUCAS Copernicus 2018: Earth-observation-relevant in situ data on land cover and use throughout the European Union. Earth System Science Data 13, 1119–1133. https://doi.org/10.5194/ESSD-13-1119-2021

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G., van der Velde, M., 2020. Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union. Scientific Data 7, 1–15. https://doi.org/10.1038/s41597-020-00675-z

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., van der Velde, M., 2022. Monitoring crop phenology with street-level imagery using computer vision. Computers and Electronics in Agriculture 196, 106866. https://doi.org/10.1016/J.COMPAG.2022.106866

d'Andrimont, R., Yordanov, M., Sedano, F., Verhegghen, A., Strobl, P., Zachariadis, S., Camilleri, F., Palmieri, A., Eiselt, B., Rubio Iglesias, J.M., van der Velde, M., 2024. Advances in LUCAS Copernicus 2022: enhancing Earth observations with comprehensive in situ data on EU land cover and use. Earth System Science Data 16, 5723–5735. https://doi.org/10.5194/essd-16-5723-2024

De Bruin, S., Bregt, A., Van De Ven, M., 2001. Assessing fitness for use: the expected value of spatial data sets. International Journal of Geographical Information Science 15, 457–471. https://doi.org/10.1080/13658810110053116

de Lima, R.A.F., Phillips, O.L., Duque, A., Tello, J.S., Davies, S.J., de Oliveira, A.A., Muller, S., Honorio Coronado, E.N., Vilanova, E., Cuni-Sanchez, A., Baker, T.R., Ryan, C.M., Malizia, A., Lewis, S.L., ter Steege, H., Ferreira, J., Marimon, B.S., Luu, H.T., Imani, G., Arroyo, L., Blundo, C., Kenfack, D., Sainge, M.N., Sonké, B., Vásquez, R., 2022. Making forest data fair and open. Nature Ecology & Evolution 2022 6:6 6, 656–658. https://doi.org/10.1038/s41559-022-01738-7

De Luca, G., Silva, J.M.N., Cerasoli, S., Araújo, J., Campos, J., Di Fazio, S., Modica, G., 2019. Object-Based Land Cover Classification of Cork Oak Woodlands using UAV Imagery and Orfeo ToolBox. Remote Sensing 2019, Vol. 11, Page 1238 11, 1238. https://doi.org/10.3390/RS11101238

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P.S., Traore, S.S., Koetz, B., 2019. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sensing of Environment 221, 551–568. https://doi.org/10.1016/J.RSE.2018.11.007

Defourny, P., Lamarche, C., Bontemps, S., De Maet, T., D Van Bogaert, E., Moreau, I., Brockmann, C., Boettcher, M., Kirches, G., Wevers, J., Santoro, M., 2017. Land Cover CCI Produce User Guide, version 2.0.

Defourny, P., Lamarche, C., Marissiaux, Q., 2020. Product Quality Assessment Report. ICDR Land Cover 2016-2020.

Defourny, P., Mayaux, P., Herold, M., Bontemps, S., 2012. Global land-cover map validation experiences remote sensing of land use and land cover. CRC Press 207–224.
Defourny, P., Schouten, L., Bartalev, S., Bontemps, S., Caccetta, P., De Wit, A.J.W., Di Bella, C., Gérard, B., Giri, C., Gond, V., Hazeu, G.W., Heinimann, A., Herold, M., Knoops J, Jaffrain, G., Latifovic, R., Lin, H., Mayaux, P., Mücher, C.A., Nonguierma, A., Stibig, H.J., Van Bogaert, E., Vancutsem, C., Bicheron, P., Leroy, M., Arino, O., 2009. Accuracy Assessment of a 300 m Global Land Cover Map: The GlobCover Experience, in: Sustaining the Millennium Development Goals. Proceedings of the 33rd International Symposium on Remote Sensing of Environment.
DeFries, R.S., Los, S.O., 1999. Implications of Land-Cover Misclassification for Parameter Estimates in Global Land-Surface Models: An Example from the Simple Biosphere Model (SiB2). Photogrammetric Engineering & Remote Sensing 65, 1083–1088.
Dewitz, J., 2021. National Land Cover Database (NLCD) 2019 Products (ver. 2.0, June 2021): USGS Science Data Catalog [Data set].
Di Gregorio, A., 2005. Land cover classification system: classification concepts and user manual: LCCS (Vol. 2). Food and Agriculture Organization of the United Nations, Rome.
Di Gregorio, A., Leonardi, U., 2016. Land Cover Classification System Software version 3 – User Manual. Food and Agriculture Organization of the United Nations, Rome, Italy.
DiMiceli, C., Sohlberg, R., Townshend, J., 2022. MODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250m SIN Grid V061 [Dataset].
Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. Science of Remote Sensing 1, 100002. https://doi.org/10.1016/J.SRS.2020.100002
Dwivedi, D., Santos, A.L.D., Barnard, M.A., Crimmins, T.M., Malhotra, A., Rod, K.A., Aho, K.S., Bell, S.M., Bomfim, B., Brearley, F.Q., Cadillo-Quiroz, H., Chen, J., Gough, C.M., Graham, E.B., Hakkenberg, C.R., Haygood, L., Koren, G., Lilleskov, E.A., Meredith, L.K., Naeher, S., Nickerson, Z.L., Pourret, O., Song, H.S., Stahl, M., Taş, N., Vargas, R., Weintraub-Leff, S., 2022. Biogeosciences Perspectives on Integrated, Coordinated, Open, Networked (ICON) Science. Earth and Space Science 9, e2021EA002119. https://doi.org/10.1029/2021EA002119
Dwyer, J.L., Roy, D.P., Sauer, B., Jenkerson, C.B., Zhang, H.K., Lymburner, L., 2018. Analysis Ready Data: Enabling Analysis of the Landsat Archive. Remote Sensing 2018, Vol. 10, Page 1363 10, 1363. https://doi.org/10.3390/RS10091363
Ebrahimy, H., Mirbagheri, B., Matkan, A.A., Azadbakht, M., 2021. Per-pixel land cover accuracy prediction: A random forest-based method with limited reference sample data. ISPRS Journal of Photogrammetry and Remote Sensing 172, 17–27. https://doi.org/10.1016/J.ISPRSJPRS.2020.11.024
ECMWF, 2020a. C3S Product Quality Assurance Document (PQAD) – ICDR Land Cover 2016-2020.
ECMWF, 2020b. C3S Product Quality Assessment Report (PQAR) – ICDR Land Cover 2016-2020.
Ellixson, A., Griffin, T., 2016. Farm Data: Ownership and Protections. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.2839811
Emmert-Streib, F., Moutari, S., Dehmer, M., 2019. A comprehensive survey of error measures for evaluating binary decision making in data science. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1303. https://doi.org/10.1002/WIDM.1303
Estes, J.E., Simonett, D.S., 1975. Fundamentals of Image Interpretation, in: Reeves, R.G. (Ed.), Manual of Remote Sensing. American Society of Photogrammetry, Falls Church, VA, pp. 869–1076.

European Commission, J., 2022. EUCROPMAP 2022 [Dataset].
FAO, 2020. Global Forest Resources Assessment 2020. Food and Agriculture Organization of the United Nations, Rome.
Feng, M., Li, X., 2020. Land cover mapping toward finer scales. Science Bulletin 65, 1604–1606. https://doi.org/10.1016/j.scib.2020.06.014
Feng, M., Sexton, J.O., Channan, S., Townshend, J.R., 2016a. A global, high-resolution (30-m) inland water body dataset for 2000: first results of a topographic–spectral classification algorithm. International Journal of Digital Earth 9, 113–133. https://doi.org/10.1080/17538947.2015.1026420
Feng, M., Sexton, J.O., Huang, C., Anand, A., Channan, S., Song, X.P., Song, D.X., Kim, D.H., Noojipady, P., Townshend, J.R., 2016b. Earth science data records of global forest cover and change: Assessment of accuracy in 1990, 2000, and 2005 epochs. Remote Sensing of Environment 184, 73–85. https://doi.org/10.1016/j.rse.2016.06.012
Ferreira, K.R., Queiroz, G.R., Camara, G., Souza, R.C.M., Vinhas, L., Marujo, R.F.B., Simoes, R.E.O., Noronha, C.A.F., Costa, R.W., Arcanjo, J.S., Gomes, V.C.F., Zaglia, M.C., 2020. Using Remote Sensing Images and Cloud Services on Aws to Improve Land Use and Cover Monitoring. 2020 IEEE Latin American GRSS and ISPRS Remote Sensing Conference, LAGIRS 2020 - Proceedings 558–562. https://doi.org/10.1109/LAGIRS48042.2020.9165649
Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global consequences of land use. Science 309, 570–574. https://doi.org/10.1126/SCIENCE.1111772/SUPPL_FILE/FOLEY_SOM.PDF
Fonte, C.C., Bastin, L., See, L., Foody, G., Lupia, F., 2015. Usability of VGI for validation of land cover maps. International Journal of Geographical Information Science 29, 1269–1291. https://doi.org/10.1080/13658816.2015.1018266
Foody, G., 2008. Harshness in image classification accuracy assessment. International Journal of Remote Sensing 29, 3137–3158. https://doi.org/10.1080/01431160701442120
Foody, G.M., 2024. Ground Truth in Classification Accuracy Assessment: Myth and Reality. Geomatics 2024, Vol. 4, Pages 81-90 4, 81–90. https://doi.org/10.3390/GEOMATICS4010005
Foody, G.M., 2023. Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. PLOS ONE 18, e0291908. https://doi.org/10.1371/JOURNAL.PONE.0291908
Foody, G.M., 2022. Global and Local Assessment of Image Classification Quality on an Overall and Per-Class Basis without Ground Reference Data. Remote Sensing 2022, Vol. 14, Page 5380 14, 5380. https://doi.org/10.3390/RS14215380
Foody, G.M., 2021. Impacts of ignorance on the accuracy of image classification and thematic mapping. Remote Sensing of Environment 259, 112367. https://doi.org/10.1016/J.RSE.2021.112367
Foody, G.M., 2020. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. Remote Sensing of Environment 239, 111630. https://doi.org/10.1016/J.RSE.2019.111630
Foody, G.M., 2014. Rating crowdsourced annotations: evaluating contributions of variable quality and completeness. International Journal of Digital Earth 7, 650–670. https://doi.org/10.1080/17538947.2013.839008
Foody, G.M., 2013. Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. Remote Sensing Letters 4, 783–792. https://doi.org/10.1080/2150704X.2013.798708

Foody, G.M., 2012. Latent class modeling for site- and non-site-specific classification accuracy assessment without ground data. IEEE Transactions on Geoscience and Remote Sensing 50, 2827–2838. https://doi.org/10.1109/TGRS.2011.2174156
Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. Remote Sensing of Environment 114, 2271–2285. https://doi.org/10.1016/J.RSE.2010.05.003
Foody, G.M., 2009. The impact of imperfect ground reference data on the accuracy of land cover change estimation. International Journal of Remote Sensing 30, 3275–3281. https://doi.org/10.1080/01431160902755346
Foody, G.M., 2005. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. International Journal of Remote Sensing 26, 1217–1228. https://doi.org/10.1080/01431160512331326521
Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80, 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4
Foody, G.M., 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. International Journal of Remote Sensing 17, 1317–1340. https://doi.org/10.1080/01431169608948706
Foody, G.M., Campbell, N.A., Trodd, N.M., 1992. Derivation and Applications of Probabilistic Measures of Class Membership from the Maximum-Likelihood Classification. Photogrammetric Engineering & Remote Sensing 58, 1335–1341.
Fortin, J.A., Cardille, J.A., Perez, E., 2020. Multi-sensor detection of forest-cover change across 45 years in Mato Grosso, Brazil. Remote Sensing of Environment 238, 111266. https://doi.org/10.1016/J.RSE.2019.111266
Frantz, D., 2019. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. Remote Sensing 11, 1124. https://doi.org/10.3390/rs11091124
Friedl, M.A., Brodley, C.E., Strahler, A.H., 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Transactions on Geoscience and Remote Sensing 37, 969–977. https://doi.org/10.1109/36.752215
Friedl, M.A., Sulla-Menashe, D., 2022. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061 [Data set].
Friedl, M.A., Woodcock, C.E., Olofsson, P., Zhu, Z., Loveland, T., Stanimirova, R., Arevalo, P., Bullock, E., Hu, K.T., Zhang, Y., Turlej, K., Tarrio, K., McAvoy, K., Gorelick, N., Wang, J.A., Barber, C.P., Souza, C., 2022. Medium Spatial Resolution Mapping of Global Land Cover and Land Cover Change Across Multiple Decades From Landsat. Frontiers in Remote Sensing 3, 894571. https://doi.org/10.3389/FRSEN.2022.894571/BIBTEX
Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Bakker, D.C.E., Hauck, J., Landschützer, P., Le Quéré, C., Luijkx, I.T., Peters, G.P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J.G., Ciais, P., Jackson, R.B., Alin, S.R., Anthoni, P., Barbero, L., Bates, N.R., Becker, M., Bellouin, N., Decharme, B., Bopp, L., Brasika, I.B.M., Cadule, P., Chamberlain, M.A., Chandra, N., Chau, T.T.T., Chevallier, F., Chini, L.P., Cronin, M., Dou, X., Enyo, K., Evans, W., Falk, S., Feely, R.A., Feng, L., Ford, D.J., Gasser, T., Ghattas, J., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Heinke, J., Houghton, R.A., Hurtt, G.C., Iida, Y., Ilyina, T., Jacobson, A.R., Jain, A., Jarníková, T., Jersild, A., Jiang, F., Jin, Z., Joos, F., Kato, E., Keeling, R.F., Kennedy, D., Goldewijk, K.K., Knauer, J., Korsbakken, J.I., Körtzinger, A., Lan, X., Lefèvre, N., Li, H., Liu, J., Liu, Z., Ma, L., Marland, G., Mayot, N., McGuire, P.C., McKinley, G.A., Meyer, G., Morgan, E.J., Munro, D.R., Nakaoka, S.I., Niwa, Y., O'Brien, K.M., Olsen, A., Omar, A.M., Ono, T., Paulsen, M., Pierrot, D., Pocock, K., Poulter, B., Powis, C.M., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T.M., Schwinger, J., Séférian, R., Smallman, T.L., Smith, S.M., Sospedra-Alfonso, R., Sun, Q., Sutton, A.J., Sweeney, C., Takao, S., Tans, P.P., Tian, H.,

Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G.R., van Ooijen, E., Wanninkhof, R., Watanabe, M., Wimart-Rousseau, C., Yang, D., Yang, X., Yuan, W., Yue, X., Zaehle, S., Zeng, J., Zheng, B., 2023. Global Carbon Budget 2023. Earth System Science Data 15, 5301–5369. https://doi.org/10.5194/ESSD-15-5301-2023

Fritz, S., Bartholomé, E., Belward, A., Hartley, A., Stibig, H.-J., Eva, H., Mayaux, P., Bartalev, S., Latifovic, R., Kolmert, S., Roy, P.S., Agrawal, S., Bingfang, W., Wenting, X., Ledwith, M., Pekel, J.-F., Giri, C., Mücher, S., De Badts, E., Tateishi, R., Champeaux, J.-L., Defourny, P., 2003. Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version). Office for Official Publications of the European Communities, Luxembourg.

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., van der Velde, M., Kraxner, F., Obersteiner, M., 2012. Geo-Wiki: An online platform for improving global land cover. Environmental Modelling & Software 31, 110–123. https://doi.org/10.1016/J.ENVSOFT.2011.11.015

Fritz, S., See, L., Bayas, J.C.L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu, B., Yan, N., You, L., Gilliams, S., Mücher, S., Tetrault, R., Moorthy, I., McCallum, I., 2019. A comparison of global agricultural monitoring systems and current gaps. Agricultural Systems 168, 258–272. https://doi.org/10.1016/J.AGSY.2018.05.010

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., Van Der Velde, M., Boettcher, H., Havlík, P., Achard, F., 2011. Highlighting continued uncertainty in global land cover maps for the user community. Environmental Research Letters 6, 044005. https://doi.org/10.1088/1748-9326/6/4/044005

Gallego, J., Delincé, J., 2010. The European Land Use and Cover Area-Frame Statistical Survey, in: Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (Eds.), Agricultural Survey Methods. John Wiley & Sons, Ltd, Chichester, pp. 149–168. https://doi.org/10.1002/9780470665480.CH10

Gao, H., Jia, G., Fu, Y., 2020. Identifying and Quantifying Pixel-Level Uncertainty among Major Satellite Derived Global Land Cover Products. Journal of Meteorological Research 34, 806–821. https://doi.org/10.1007/S13351-020-9183-X/METRICS

Gao, P., Cushman, S.A., Liu, G., Ye, S., Shen, S., Cheng, C., 2019. FracL: A Tool for Characterizing the Fractality of Landscape Gradients from a New Perspective. ISPRS International Journal of Geo-Information 2019, Vol. 8, Page 466 8, 466. https://doi.org/10.3390/IJGI8100466

García-Álvarez, D., Lara Hinojosa, J., Jurado Pérez, F.J., Quintero Villaraso, J., 2022. Global General Land Use Cover Datasets with a Time Series of Maps, in: García-Álvarez, David, Camacho Olmedo, M.T., Paegelow, M., Mas, J.F. (Eds.), Land Use Cover Datasets and Validation Tools: Validation Practices with QGIS. Springer International Publishing, Cham, Switzerland, pp. 287–311.

GCOS, 2011. Systematic observation requirements for satellite-based data products for climate: Supplemental details to the satellite-based component of the Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. World Meteorological Organization.

GFOI, 2020. Integration of Remote-Sensing and Ground-Based Observations for Estimation of Emissions and Removals of Greenhouse Gases in Forests: Methods and Guidance from the Global Forest Observations Initiative. Edition 3.0.

GFOI, 2016. Integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests: Methods and Guidance from the Global Forest Observations Initiative. Edition 2.0. Food and Agriculture Organization, Rome.

Giglio, L., Schroeder, W., Justice, C.O., 2016. The collection 6 MODIS active fire detection algorithm and fire products. Remote Sensing of Environment 178, 31–41. https://doi.org/10.1016/J.RSE.2016.02.054

Giri, C., Ochieng, E., Tieszen, L.L., Zhu, Z., Singh, A., Loveland, T., Masek, J., Duke, N., 2011. Status and distribution of mangrove forests of the world using earth observation satellite data. Global Ecology and Biogeography 20, 154–159. https://doi.org/10.1111/J.1466-8238.2010.00584.X

Giri, C., Zhu, Z., Reed, B., 2005. A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets. Remote Sensing of Environment 94, 123–132. https://doi.org/10.1016/J.RSE.2004.09.005

Glushkov, I., Zhuravleva, I., McCarty, J.L., Komarova, A., Drozdovsky, A., Drozdovskaya, M., Lupachik, V., Yaroshenko, A., Stehman, S.V., Prishchepov, A.V., 2021. Spring fires in Russia: results from participatory burned area mapping with Sentinel-2 imagery. Environmental Research Letters 16, 125005. https://doi.org/10.1088/1748-9326/AC3287

Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, Wenyu, Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, Weijia, Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q., Song, L., 2019. 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. Science bulletin 64, 370–373. https://doi.org/10.1016/J.SCIB.2019.03.002

Gong, P., Wang, J., Yu, L., Zhao, Yongchao, Zhao, Yuanyuan, Liang, L., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q., Hu, L., Yao, W., Zhang, Han, Zhu, P., Zhao, Z., Zhang, Haiying, Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A., Guo, J., Yu, Liang, Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, Jin, Chen, Jun, 2013. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. International Journal of Remote Sensing 34, 2607–2654. https://doi.org/10.1080/01431161.2012.748992

Goodchild, M.F., 2007. Citizens as sensors: The world of volunteered geography. GeoJournal 69, 211–221. https://doi.org/10.1007/S10708-007-9111-Y/FIGURES/8

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Goryl, P., Fox, N., Donlon, C., Castracane, P., 2023. Fiducial Reference Measurements (FRMs): What Are They? Remote Sensing 15, 5017. https://doi.org/10.3390/rs15205017

Group on Earth Observations, 2018. Joint Experiment for Crop Assessment and Monitoring (JECAM) Guidelines for cropland and crop type definition and field data collection.

Halladin-Dabrowska, A., Kania, A., Kopeć, D., 2019. The t-SNE Algorithm as a Tool to Improve the Quality of Reference Data Used in Accurate Mapping of Heterogeneous Non-Forest Vegetation. Remote Sensing 2020, Vol. 12, Page 39 12, 39. https://doi.org/10.3390/RS12010039

Hancock, S., McGrath, C., Lowe, C., Davenport, I., Woodhouse, I., 2021. Requirements for a global lidar system: spaceborne lidar with wall-to-wall coverage. Royal Society Open Science 8. https://doi.org/10.1098/RSOS.211166

Hansen, M.C., Krylov, A., Tyukavina, A., Potapov, P.V., Turubanova, S., Zutta, B., Ifo, S., Margono, B., Stolle, F., Moore, R., 2016. Humid tropical forest disturbance alerts using Landsat data. Environmental Research Letters 11, 34008. https://doi.org/10.1088/1748-9326/11/3/034008

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., Kommareddy, A., Egorov, A., Chini, L.,

Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21-st century forest cover change. Science (New York, N.Y.) 342, 850–853.
Hansen, M.C., Potapov, P.V., Pickens, A.H., Tyukavina, A., Hernandez-Serna, A., Zalles, V., Turubanova, S., Kommareddy, I., Stehman, S.V., Song, X.P., Kommareddy, A., 2022. Global land use extent and dispersion within natural land cover using Landsat data. Environmental Research Letters 17. https://doi.org/10.1088/1748-9326/ac46ec
Hay, A.M., 1979. Sampling designs to test land-use map accuracy. Photogrammetric Engineering and Remote Sensing 45, 529–533.
Herold, M., Mayaux, P., Woodcock, C.E., Baccini, A., Schmullius, C., 2008. Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. Remote Sensing of Environment 112, 2538–2556. https://doi.org/10.1016/j.rse.2007.11.013
Herold, M., See, L., Tsendbazar, N.E., Fritz, S., 2016. Towards an Integrated Global Land Cover Monitoring and Mapping System. Remote Sensing 2016, Vol. 8, Page 1036 8, 1036. https://doi.org/10.3390/RS8121036
Herold, M., Woodcock, C., Cihlar, J., Wulder, M., Arino, O., Achard, F., Hansen, M., Holsson, H., Schmulllius, C., Brady, M., Di Gregorio, A., Latham, J., Sessa, R., 2009. Assessment of the Status of the Development of the Standards for the Terrestrial Essential Climate Variables: T9 Land Cover. Global Terrestrial Observing System, Rome.
Horning, N., Fleishman, E., Ersts, P.J., Fogarty, F.A., Wohlfeil Zillig, M., 2020. Mapping of land cover with open-source software and ultra-high-resolution imagery acquired with unmanned aerial vehicles. Remote Sensing in Ecology and Conservation 6, 487–497. https://doi.org/10.1002/RSE2.144/
Howe, J., 2006. The Rise of Crowdsourcing. Wired Magazine 14, 1–4.
IPCC, 2006. 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Prepared by the National Greenhouse Gas Inventories Programme. IGES, Japan.
Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next Landsat satellite: The Landsat Data Continuity Mission. Remote Sensing of Environment 122, 11–21. https://doi.org/10.1016/J.RSE.2011.08.026
Iwao, K., Nishida, K., Kinoshita, T., Yamagata, Y., 2006. Validating land cover maps with Degree Confluence Project information. Geophysical Research Letters 33, 23404. https://doi.org/10.1029/2006GL027768
Jansen, L.J.M., Groom, G., Carrai, G., 2008. Land-cover harmonisation and semantic similarity: some methodological issues. Journal of Land Use Science 3, 131–160. https://doi.org/10.1080/17474230802332076
JCGM, 2012. International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM), 3rd edition. JCGM 200: 2012. https://doi.org/10.59161/JCGM200-2012
Jenice, A.R., Raimond, K., 2015. A review on availability of remote sensing data. Proceedings - 2015 IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development, TIAR 2015 150–155. https://doi.org/10.1109/TIAR.2015.7358548
Johnson, A.C., Sumpter, J.P., 2016. Are we going about chemical risk assessment for the aquatic environment the wrong way? Environmental Toxicology and Chemistry 35, 1609–1616. https://doi.org/10.1002/ETC.3441
Johnson, D.M., 2013. A 2010 map estimate of annually tilled cropland within the conterminous United States. Agricultural Systems 114, 95–105. https://doi.org/10.1016/J.AGSY.2012.08.004
Johnson, E.W., Ross, J., 2008. Quantifying error in aerial survey data. Australian Forestry 71, 216–222. https://doi.org/10.1080/00049158.2008.10675038

Jonckheere, I., Hamilton, R., Michel, J.M., Donegan, E., 2024. Good practices in sample-based area estimation. White paper. Food and Agriculture Organization of the United Nations, Rome.
Joyce, K.E., Anderson, K., Bartolo, R.E., 2021. Of Course We Fly Unmanned—We're Women! Drones 2021, Vol. 5, Page 21 5, 21. https://doi.org/10.3390/DRONES5010021
Jumaat, N.F.H., Ahmad, B., Dutsenwai, H.S., 2018. Land cover change mapping using high resolution satellites and unmanned aerial vehicle. IOP Conference Series: Earth and Environmental Science 169, 012076. https://doi.org/10.1088/1755-1315/169/1/012076
Justice, C., Belward, A., Morisette, J., Lewis, P., Privette, J., Baret, F., 2000. Developments in the "validation" of satellite sensor products for the study of the land surface. International Journal of Remote Sensing 21, 3383–3390. https://doi.org/10.1080/014311600750020000
Justice, C.O., Giglio, L., Korontzi, S., Owens, J., Morisette, J.T., Roy, D., Descloitres, J., Alleaume, S., Petitcolin, F., Kaufman, Y., 2002. The MODIS fire products. Remote Sensing of Environment 83, 244–262. https://doi.org/10.1016/S0034-4257(02)00076-7
Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J.C., Mathis, M., Brumby, S.P., 2021. Global Land Use/Land Cover with Sentinel 2 and deep learning, in: International Geoscience and Remote Sensing Symposium (IGARSS). Institute of Electrical and Electronics Engineers Inc., pp. 4704–4707. https://doi.org/10.1109/IGARSS47720.2021.9553499
Karydas, C.G., Gitas, I.Z., Kuntz, S., Minakou, C., 2015. Use of LUCAS LC Point Database for Validating Country-Scale Land Cover Maps. Remote Sensing 2015, Vol. 7, Pages 5012-5041 7, 5012–5041. https://doi.org/10.3390/RS70505012
Khan, A., Hansen, M.C., Potapov, P., Stehman, S.V., Chatta, A.A., 2016. Landsat-based wheat mapping in the heterogeneous cropping system of Punjab, Pakistan. International Journal of Remote Sensing 37, 1391–1410. https://doi.org/10.1080/01431161.2016.1151572
Khatami, R., Mountrakis, G., Stehman, S.V., 2017. Mapping per-pixel predicted accuracy of classified remote sensing images. Remote Sensing of Environment 191, 156–167. https://doi.org/10.1016/J.RSE.2017.01.025
Klaschka, J., Reiczigel, J., 2021. On matching confidence intervals and tests for some discrete distributions: methodological and computational aspects. Computational Statistics 36, 1775–1790. https://doi.org/10.1007/S00180-020-00986-0/FIGURES/3
Koren, G., Ferrara, V., Timmins, M., Morrison, M.A., 2022. Global Environmental Change Perspectives on Integrated, Coordinated, Open, and Networked (ICON) Science. Earth and Space Science 9, e2022EA002231. https://doi.org/10.1029/2022EA002231
Kosztra, B., Büttner, G., Hazeu, G., Arnold, 2019. Updated CLC illustrated nomenclature guidelines. European Environmental Agency, Wien, Austria.
Krylov, A., Steininger, M.K., Hansen, M.C., Potapov, P.V., Stehman, S.V., Gost, A., Noel, J., Talero Ramirez, Y., Tyukavina, A., Di Bella, C.M., Ellis, E.A., Ellis, P., 2018. Contrasting tree-cover loss and subsequent land cover in two neotropical forest regions: sample-based assessment of the Mexican Yucatán and Argentine Chaco. Journal of Land Use Science 13, 549–564. https://doi.org/10.1080/1747423X.2019.1569169
Lamarche, C., Bontemps, S., Marissiaux, Q., Defourny, P., Arino, O., 2021. Towards a Multi-Level Sampling Scheme for Land Cover and Land Cover Change Validation. Lessons Learned from the Land Cover Climate Change Initiative. International Geoscience and Remote Sensing Symposium (IGARSS) 1986–1989. https://doi.org/10.1109/IGARSS47720.2021.9553898
Lamarche, C., Santoro, M., Bontemps, S., d'Andrimont, R., Radoux, J., Giustarini, L., Brockmann, C., Wevers, J., Defourny, P., Arino, O., 2017. Compilation and Validation of SAR and Optical Data Products for a Complete and Global Map of Inland/Ocean Water

Tailored to the Climate Modeling Community. Remote Sensing 2017, Vol. 9, Page 36 9, 36. https://doi.org/10.3390/RS9010036
Laso Bayas, J.C., Lesiv, M., Waldner, F., Schucknecht, A., Duerauer, M., See, L., Fritz, S., Fraisl, D., Moorthy, I., McCallum, I., Perger, C., Danylo, O., Defourny, P., Gallego, J., Gilliams, S., Akhtar, I.U.H., Baishya, S.J., Baruah, M., Bungnamei, K., Campos, A., Changkakati, T., Cipriani, A., Das, Krishna, Das, Keemee, Das, I., Davis, K.F., Hazarika, P., Johnson, B.A., Malek, Z., Molinari, M.E., Panging, K., Pawe, C.K., Pérez-Hoyos, A., Sahariah, P.K., Sahariah, D., Saikia, A., Saikia, M., Schlesinger, P., Seidacaru, E., Singha, K., Wilson, J.W., 2017. A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform. Scientific Data 2017 4:1 4, 1–10. https://doi.org/10.1038/sdata.2017.136
Laso, F.J., Benítez, F.L., Rivas-Torres, G., Sampedro, C., Arce-Nazario, J., 2020. Land Cover Classification of Complex Agroecosystems in the Non-Protected Highlands of the Galapagos Islands. Remote Sensing 12. https://doi.org/10.3390/RS12010065
Latifovic, R., Zhu, Z.L., Cihlar, J., Giri, C., Olthof, I., 2004. Land cover mapping of North and Central America—Global Land Cover 2000. Remote Sensing of Environment 89, 116–127. https://doi.org/10.1016/J.RSE.2003.11.002
Lausch, A., Blaschke, T., Haase, D., Herzog, F., Syrbe, R.U., Tischendorf, L., Walz, U., 2015. Understanding and quantifying landscape structure – A review on relevant process characteristics, data models and landscape metrics. Ecological Modelling 295, 31–41. https://doi.org/10.1016/J.ECOLMODEL.2014.08.018
Lazzeri, G., Frodella, W., Rossi, G., Moretti, S., 2021. Multitemporal Mapping of Post-Fire Land Cover Using Multiplatform PRISMA Hyperspectral and Sentinel-UAV Multispectral Data: Insights from Case Studies in Portugal and Italy. Sensors 2021, Vol. 21, Page 3982 21, 3982. https://doi.org/10.3390/S21123982
Lesiv, M., Schepaschenko, D., Buchhorn, M., See, L., Dürauer, M., Georgieva, I., Jung, M., Hofhansl, F., Schulze, K., Bilous, A., Blyshchyk, V., Mukhortova, L., Brenes, C.L.M., Krivobokov, L., Ntie, S., Tsogt, K., Pietsch, S.A., Tikhonova, E., Kim, M., Di Fulvio, F., Su, Y.F., Zadorozhniuk, R., Sirbu, F.S., Panging, K., Bilous, S., Kovalevskii, S.B., Kraxner, F., Rabia, A.H., Vasylyshyn, R., Ahmed, R., Diachuk, P., Kovalevskyi, S.S., Bungnamei, K., Bordoloi, K., Churilov, A., Vasylyshyn, O., Sahariah, D., Tertyshnyi, A.P., Saikia, A., Malek, Ž., Singha, K., Feshchenko, R., Prestele, R., Akhtar, I. ul H., Sharma, K., Domashovets, G., Spawn-Lee, S.A., Blyshchyk, O., Slyva, O., Ilkiv, M., Melnyk, O., Sliusarchuk, V., Karpuk, A., Terentiev, A., Bilous, V., Blyshchyk, K., Bilous, M., Bogovyk, N., Blyshchyk, I., Bartalev, S., Yatskov, M., Smets, B., Visconti, P., Mccallum, I., Obersteiner, M., Fritz, S., 2022. Global forest management data for 2015 at a 100 m resolution. Scientific Data 2022 9:1 9, 1–14. https://doi.org/10.1038/s41597-022-01332-3
Lesiv, M., See, L., Bayas, J.C.L., Sturn, T., Schepaschenko, D., Karner, M., Moorthy, I., McCallum, I., Fritz, S., 2018. Characterizing the Spatial and Temporal Availability of Very High Resolution Satellite Imagery in Google Earth and Microsoft Bing Maps as a Source of Reference Data. Land 2018, Vol. 7, Page 118 7, 118. https://doi.org/10.3390/LAND7040118
Li, C., Gong, P., Wang, J., Zhu, Z., Biging, G.S., Yuan, C., Hu, T., Zhang, H., Wang, Q., Li, X., Liu, X., Xu, Y., Guo, J., Liu, C., Hackman, K.O., Zhang, M., Cheng, Y., Yu, L., Yang, J., Huang, H., Clinton, N., 2017. The first all-season sample set for mapping global land cover with Landsat-8 data. Science Bulletin 62, 508–515. https://doi.org/10.1016/J.SCIB.2017.03.011
Li, C., Wang, J., Wang, L., Hu, L., Gong, P., 2014. Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery. Remote Sensing 2014, Vol. 6, Pages 964-983 6, 964–983. https://doi.org/10.3390/RS6020964

Liu, C., Frazier, P., Kumar, L., 2007. Comparative assessment of the measures of thematic classification accuracy. Remote Sensing of Environment 107, 606–616. https://doi.org/10.1016/J.RSE.2006.10.010
Liu, L., Zhang, X., Chen, X., Gao, Y., Mi, J., 2020. GLC_FCS30-2020:Global Land Cover with Fine Classification System at 30m in 2020 (v1.2) [Dataset].
Liu, X., Huang, Y., Xu, X., Li, Xuecao, Li, Xia, Ciais, P., Lin, P., Gong, K., Ziegler, A.D., Chen, A., Gong, P., Chen, J., Hu, G., Chen, Y., Wang, S., Wu, Q., Huang, K., Estes, L., Zeng, Z., 2020. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. Nature Sustainability 3, 564–570. https://doi.org/10.1038/s41893-020-0521-x
Lohr, S., 2010. Sampling: Design and Analysis. Brooks/Cole, Boston, Massachusetts.
Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J., Elith, J., 2018. A comparison of resampling methods for remote sensing classification and accuracy assessment. Remote Sensing of Environment 208, 145–153. https://doi.org/10.1016/J.RSE.2018.02.026
Mandlburger, G., Lehner, H., Pfeifer, N., 2019. A Comparison of Single Photon and Full Waveform LIDAR. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2-W5, 397–404. https://doi.org/10.5194/isprs-annals-IV-2-W5-397-2019
Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., Strano, E., 2020. Outlining where humans live, the World Settlement Footprint 2015. Scientific Data 2020 7:1 7, 1–14. https://doi.org/10.1038/s41597-020-00580-5
Maselli, F., Conese, C., Petkov, L., 1994. Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. ISPRS Journal of Photogrammetry and Remote Sensing 49, 13–20. https://doi.org/10.1016/0924-2716(94)90062-0
Masiliunas, D., Tsendbazar, N.-E., Herold, M., Verbesselt, J., Yang, W., Chen, X., Wang, C., Cao, R., Zhu, X., Shen, M., 2021. BFAST Lite: A Lightweight Break Detection Method for Time Series Analysis. Remote Sensing 2021, Vol. 13, Page 3308 13, 3308. https://doi.org/10.3390/RS13163308
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., Kopin, Y., Roy, P.S., 2006. Validation of the global land cover 2000 map. IEEE Transactions on Geoscience and Remote Sensing 44, 1728–1737. https://doi.org/10.1109/TGRS.2006.864370
McCallum, I., Obersteiner, M., Nilsson, S., Shvidenko, A., 2006. A spatial comparison of four satellite derived 1 km global land cover datasets. International Journal of Applied Earth Observation and Geoinformation 8, 246–255. https://doi.org/10.1016/J.JAG.2005.12.002
McRoberts, R.E., 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sensing of Environment 114, 1017–1025. https://doi.org/10.1016/J.RSE.2009.12.013
McRoberts, R.E., 2006. A model-based approach to estimating forest area. Remote Sensing of Environment 103, 56–66. https://doi.org/10.1016/J.RSE.2006.03.005
McRoberts, R.E., Næsset, E., Saatchi, S., Quegan, S., 2022. Statistically rigorous, model-based inferences from maps. Remote Sensing of Environment 279, 113028. https://doi.org/10.1016/J.RSE.2022.113028
McRoberts, R.E., Stehman, S.V., Liknes, G.C., Næsset, E., Sannier, C., Walters, B.F., 2018. The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions. ISPRS Journal of Photogrammetry and Remote Sensing 142, 292–300. https://doi.org/10.1016/j.isprsjprs.2018.06.002
Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nature Communications 2022 13:1 13, 1–4. https://doi.org/10.1038/s41467-022-29838-9

Morisette, J.T., Privette, J.L., Justice, C.O., 2002. A framework for the validation of MODIS Land products. Remote Sensing of Environment 83, 77–96. https://doi.org/10.1016/S0034-4257(02)00088-3

Nagatani, I., Hayashi, M., Watanabe, M., Tadano, T., Watanabe, T., Koyama, C., Shimada, M., 2018. Forest early warning system using ALOS-2/PALSAR-2 SCanSAR data (JJ-FAST). International Geoscience and Remote Sensing Symposium (IGARSS) 2018-July, 4181–4184. https://doi.org/10.1109/IGARSS.2018.8517431

Nakalembe, C., Becker-Reshef, I., Bonifacio, R., Hu, G., Humber, M.L., Justice, C.J., Keniston, J., Mwangi, K., Rembold, F., Shukla, S., Urbano, F., Whitcraft, A.K., Li, Y., Zappacosta, M., Jarvis, I., Sanchez, A., 2021. A review of satellite-based global agricultural monitoring systems available for Africa. Global Food Security 29, 100543. https://doi.org/10.1016/J.GFS.2021.100543

Natesan, S., Armenakis, C., Benari, G., Lee, R., 2018. Use of UAV-Borne Spectrometer for Land Cover Classification. Drones 2018, Vol. 2, Page 16 2, 16. https://doi.org/10.3390/DRONES2020016

Nex, F., Armenakis, C., Cramer, M., Cucci, D.A., Gerke, M., Honkavaara, E., Kukko, A., Persello, C., Skaloud, J., 2022. UAV in the advent of the twenties: Where we stand and what is next. ISPRS Journal of Photogrammetry and Remote Sensing 184, 215–242. https://doi.org/10.1016/J.ISPRSJPRS.2021.12.006

NICFI Data Program, 2021. User guide.

Olofsson, P., Arévalo, P., Espejo, A.B., Green, C., Lindquist, E., McRoberts, R.E., Sanz, M.J., 2020. Mitigating the effects of omission errors on area and area change estimates. Remote Sensing of Environment 236, 111492. https://doi.org/10.1016/J.RSE.2019.111492

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sensing of Environment 129, 122–131. https://doi.org/10.1016/j.rse.2012.10.031

Olofsson, P., Stehman, S.V., Woodcock, C.E., Sulla-Menashe, D., Sibley, A.M., Newell, J.D., Friedl, M.A., Herold, M., 2012. A global land-cover validation data set, part I: fundamental design principles. International Journal of Remote Sensing 33, 5768–5788. https://doi.org/10.1080/01431161.2012.674230

Park, N.W., Kyriakidis, P.C., Hong, S.Y., 2016. Spatial Estimation of Classification Accuracy Using Indicator Kriging with an Image-Derived Ambiguity Index. Remote Sensing 2016, Vol. 8, Page 320 8, 320. https://doi.org/10.3390/RS8040320

Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584

Pengra, B., Long, J., Dahal, D., Stehman, S.V., Loveland, T.R., 2015. A global reference database from very high resolution commercial satellite data and methodology for application to Landsat derived 30m continuous field tree cover data. Remote Sensing of Environment 165, 234–248. https://doi.org/10.1016/j.rse.2015.01.018

Pengra, B.W., Stehman, S.V., Horton, J.A., Dockter, D.J., Schroeder, T.A., Yang, Z., Cohen, W.B., Healey, S.P., Loveland, T.R., 2020. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. Remote Sensing of Environment 238, 111261. https://doi.org/10.1016/J.RSE.2019.111261

Pesaresi, M., Politis, P., 2023. GHS-BUILT-S R2023A - GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030) [Dataset]. https://doi.org/10.2905/9F06F36F-4B11-47EC-ABB0-4F8B7B1D72EA

Pickens, A.H., Hansen, M.C., Hancher, M., Stehman, S.V., Tyukavina, A., Potapov, P., Marroquin, B., Sherani, Z., 2020. Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. Remote Sensing of Environment 243. https://doi.org/10.1016/j.rse.2020.111792

Pickens, A.H., Hansen, M.C., Stehman, S.V., Tyukavina, A., Potapov, P., Zalles, V., Higgins, J., 2022. Global seasonal dynamics of inland open water and ice. Remote Sensing of Environment 272. https://doi.org/10.1016/j.rse.2022.112963

Pickering, J., Stehman, S.V., Tyukavina, A., Potapov, P., Watt, P., Jantz, S.M., Bholanath, P., Hansen, M.C., 2019. Quantifying the trade-off between cost and precision in estimating area of forest loss and degradation using probability sampling in Guyana. Remote Sensing of Environment 221, 122–135. https://doi.org/10.1016/j.rse.2018.11.018

Pickering, J., Tyukavina, A., Khan, A., Potapov, P., Adusei, B., Hansen, M.C., Lima, A., Deng, C., 2021. Using Multi-Resolution Satellite Data to Quantify Land Dynamics: Applications of PlanetScope Imagery for Cropland and Tree-Cover Loss Area Estimation. https://doi.org/10.3390/rs

Pontius, R.G., Krithivasan, R., Sauls, L., Yan, Y., Zhang, Y., 2017. Methods to summarize change among land categories across time intervals. Journal of Land Use Science 12, 218–230. https://doi.org/10.1080/1747423X.2017.1338768

Pontius, R.G., Lippitt, C.D., 2006. Can Error Explain Map Differences Over Time? Cartography and Geographic Information Science 33, 159–171. https://doi.org/10.1559/152304006777681706

Pontius, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32, 4407–4429. https://doi.org/10.1080/01431161.2011.552923

Potapov, P., Hansen, M.C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., Ying, Q., 2020. Landsat analysis ready data for global land cover and land cover change mapping. Remote Sensing 12. https://doi.org/10.3390/rs12030426

Potapov, P., Hansen, M.C., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S., Zalles, V., Li, X., Khan, A., Stolle, F., Harris, N., Song, X.-P., Baggett, A., Kommareddy, I., Kommareddy, A., 2022a. The Global 2000-2020 Land Cover and Land Use Change Dataset Derived From the Landsat Archive: First Results. Frontiers in Remote Sensing 3. https://doi.org/10.3389/frsen.2022.856903

Potapov, P., Hansen, M.C., Turubanova, S., Tyukavina, A., Zalles, V., Song, X.P., Khan, A., 2023. Reply to: Measuring the world's cropland area. Nature Food 2023 4:1 4, 33–34. https://doi.org/10.1038/s43016-022-00668-8

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J.B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sensing of Environment 253. https://doi.org/10.1016/j.rse.2020.112165

Potapov, P., Siddiqui, B.N., Iqbal, Z., Aziz, T., Zzaman, B., Islam, A., Pickens, A., Talero, Y., Tyukavina, A., Turubanova, S., Hansen, M.C., 2017. Comprehensive monitoring of Bangladesh tree cover inside and outside of forests, 2000–2014. Environmental Research Letters 12, 104015. https://doi.org/10.1088/1748-9326/AA84BB

Potapov, P., Turubanova, S., Hansen, M.C., Tyukavina, A., Zalles, V., Khan, A., Song, X.P., Pickens, A., Shen, Q., Cortez, J., 2022b. Global maps of cropland extent and change

show accelerated cropland expansion in the twenty-first century. Nature Food 3, 19–28. https://doi.org/10.1038/s43016-021-00429-z
Potapov, P., Tyukavina, A., Turubanova, S., Talero, Y., Hernandez-Serna, A., Hansen, M.C., Saah, D., Tenneson, K., Poortinga, A., Aekakkararungroj, A., Chishtie, F., Towashiraporn, P., Bhandari, B., Aung, K.S., Nguyen, Q.H., 2019. Annual continuous fields of woody vegetation structure in the Lower Mekong region from 2000-2017 Landsat time-series. Remote Sensing of Environment 232, 111278. https://doi.org/10.1016/J.RSE.2019.111278
Potapov, P.V., Dempewolf, J., Talero, Y., Hansen, M.C., Stehman, S.V., Vargas, C., Rojas, E.J., Castillo, D., Mendoza, E., Calderón, A., Giudice, R., Malaga, N., Zutta, B.R., 2014. National satellite-based humid tropical forest change assessment in Peru in support of REDD+ implementation. Environmental Research Letters 9, 124012. https://doi.org/10.1088/1748-9326/9/12/124012
Potapov, P.V., Turubanova, S.A., Hansen, M.C., Adusei, B., Broich, M., Altstatt, A., Mane, L., Justice, C.O., 2012. Quantifying forest cover loss in Democratic Republic of the Congo, 2000-2010, with Landsat ETM+ data. Remote Sensing of Environment 122, 106–116. https://doi.org/10.1016/j.rse.2011.08.027
Potapov, P.V., Turubanova, S.A., Tyukavina, A., Krylov, A.M., McCarty, J.L., Radeloff, V.C., Hansen, M.C., 2015. Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive. Remote Sensing of Environment 159, 28–43. https://doi.org/10.1016/j.rse.2014.11.027
Pouliot, D., Latifovic, R., Fernandes, R., Olthof, I., 2009. Evaluation of annual forest disturbance monitoring using a static decision tree approach and 250 m MODIS data. Remote Sensing of Environment 113, 1749–1759. https://doi.org/10.1016/j.rse.2009.04.008
Powell, R.L., Matzke, N., De Souza, C., Clark, M., Numata, I., Hess, L.L., Roberts, D.A., Clark, M., Numata, I., Hess, L.L., Roberts, D.A., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sensing of Environment 90, 221–234. https://doi.org/10.1016/J.RSE.2003.12.007
Praveen, A., Jeganathan, C., Mondal, S., 2023. Mapping Annual Cropping Pattern from Time-Series MODIS EVI Using Parameter-Tuned Random Forest Classifier. Journal of the Indian Society of Remote Sensing 51, 983–1000. https://doi.org/10.1007/S12524-023-01676-2/FIGURES/12
Radeloff, V.C., Roy, D.P., Wulder, M.A., Anderson, M., Cook, B., Crawford, C.J., Friedl, M., Gao, F., Gorelick, N., Hansen, M., Healey, S., Hostert, P., Hulley, G., Huntington, J.L., Johnson, D.M., Neigh, C., Lyapustin, A., Lymburner, L., Pahlevan, N., Pekel, J.-F., Scambos, T.A., Schaaf, C., Strobl, P., Woodcock, C.E., Zhang, H.K., Zhu, Z., 2024. Need and vision for global medium-resolution Landsat and Sentinel-2 data products. Remote Sensing of Environment 300, 113918. https://doi.org/10.1016/j.rse.2023.113918
Radoux, J., Bogaert, P., 2020. About the Pitfall of Erroneous Validation Data in the Estimation of Confusion Matrices. Remote Sensing 2020, Vol. 12, Page 4128 12, 4128. https://doi.org/10.3390/RS12244128
Radoux, J., Bogaert, P., 2017. Good Practices for Object-Based Accuracy Assessment. Remote Sensing 2017, Vol. 9, Page 646 9, 646. https://doi.org/10.3390/RS9070646
Radoux, J., Bogaert, P., 2014. Accounting for the area of polygon sampling units for the prediction of primary accuracy assessment indices. Remote Sensing of Environment 142, 9–19. https://doi.org/10.1016/J.RSE.2013.10.030
Radoux, J., Bogaert, P., Fasbender, D., Defourny, P., 2011. Thematic accuracy assessment of geographic object-based image classification. International Journal of Geographical Information Science 25, 895–911. https://doi.org/10.1080/13658816.2010.498378

Radoux, J., Defourny, P., 2007. A quantitative assessment of boundaries in automated forest stand delineation using very high resolution imagery. Remote Sensing of Environment 110, 468–475. https://doi.org/10.1016/J.RSE.2007.02.031
Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C., Defourny, P., 2014. Automated Training Sample Extraction for Global Land Cover Mapping. Remote Sensing 2014, Vol. 6, Pages 3965-3987 6, 3965–3987. https://doi.org/10.3390/RS6053965
Radoux, J., Waldner, F., Bogaert, P., 2020. How Response Designs and Class Proportions Affect the Accuracy of Validation Data. Remote Sensing 2020, Vol. 12, Page 257 12, 257. https://doi.org/10.3390/RS12020257
Reiche, J., Hamunyela, E., Verbesselt, J., Hoekman, D., Herold, M., 2018. Improving near-real time deforestation monitoring in tropical dry forests by combining dense Sentinel-1 time series with Landsat and ALOS-2 PALSAR-2. Remote Sensing of Environment 204, 147–161. https://doi.org/10.1016/J.RSE.2017.10.034
Reiche, J., Mullissa, A., Slagter, B., Gou, Y., Tsendbazar, N.E., Odongo-Braun, C., Vollrath, A., Weisse, M.J., Stolle, F., Pickens, A., Donchyts, G., Clinton, N., Gorelick, N., Herold, M., 2021. Forest disturbance alerts for the Congo Basin using Sentinel-1. Environmental Research Letters 16, 024005. https://doi.org/10.1088/1748-9326/ABD0A8
Rice, J.A., 2007. Mathematical Statistics and Data Analysis.
Riebsame, W.E., Meyer, W.B., Turner, B.L., 1994. Modeling land use and cover as part of global environmental change. Climatic Change 28, 45–64. https://doi.org/10.1007/BF01094100/METRICS
Riemann, R., Wilson, B.T., Lister, A., Parks, S., 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. Remote Sensing of Environment 114, 2337–2352. https://doi.org/10.1016/j.rse.2010.05.010
Saah, D., Johnson, G., Ashmall, B., Tondapu, G., Tenneson, K., Patterson, M., Poortinga, A., Markert, K., Quyen, N.H., San Aung, K., Schlichting, L., Matin, M., Uddin, K., Aryal, R.R., Dilger, J., Lee Ellenburg, W., Flores-Anderson, A.I., Wiell, D., Lindquist, E., Goldstein, J., Clinton, N., Chishtie, F., 2019. Collect Earth: An online tool for systematic reference data collection in land cover and use applications. Environmental Modelling & Software 118, 166–171. https://doi.org/10.1016/J.ENVSOFT.2019.05.004
Saah, D., Tenneson, K., Poortinga, A., Nguyen, Q., Chishtie, F., Aung, K.S., Markert, K.N., Clinton, N., Anderson, E.R., Cutter, P., Goldstein, J., Housman, I.W., Bhandari, B., Potapov, P.V., Matin, M., Uddin, K., Pham, H.N., Khanal, N., Maharjan, S., Ellenberg, W.L., Bajracharya, B., Bhargava, R., Maus, P., Patterson, M., Flores-Anderson, A.I., Silverman, J., Sovann, C., Do, P.M., Nguyen, G.V., Bounthabandit, S., Aryal, R.R., Myat, S.M., Sato, K., Lindquist, E., Kono, M., Broadhead, J., Towashiraporn, P., Ganz, D., 2020. Primitives as building blocks for constructing land cover maps. International Journal of Applied Earth Observation and Geoinformation 85, 101979. https://doi.org/10.1016/J.JAG.2019.101979
Saarela, S., Holm, S., Healey, S.P., Andersen, H.E., Petersson, H., Prentius, W., Patterson, P.L., Næsset, E., Gregoire, T.G., Ståhl, G., 2018. Generalized Hierarchical Model-Based Estimation for Aboveground Biomass Assessment Using GEDI and Landsat Data. Remote Sensing 2018, Vol. 10, Page 1832 10, 1832. https://doi.org/10.3390/RS10111832
Sales, M.H.R., De Bruin, S., Souza, C., Herold, M., 2022. Land Use and Land Cover Area Estimates from Class Membership Probability of a Random Forest Classification. IEEE Transactions on Geoscience and Remote Sensing 60. https://doi.org/10.1109/TGRS.2021.3080083

Santoro, M., Wegmüller, U., 2014. Multi-temporal synthetic aperture radar metrics applied to map open water bodies. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, 3225–3238. https://doi.org/10.1109/JSTARS.2013.2289301

Sarmento, P., Carrão, H., Caetano, M., Stehman, S.V., 2009. Incorporating reference classification uncertainty into the analysis of land cover accuracy. International Journal of Remote Sensing 30, 5309–5321. https://doi.org/10.1080/01431160903130994

Särndal, C.-E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Schepaschenko, D., See, L., Lesiv, M., Bastin, J.F., Mollicone, D., Tsendbazar, N.E., Bastin, L., McCallum, I., Laso Bayas, J.C., Baklanov, A., Perger, C., Dürauer, M., Fritz, S., 2019. Recent Advances in Forest Observation with Visual Interpretation of Very High-Resolution Imagery. Surveys in Geophysics 40, 839–862. https://doi.org/10.1007/S10712-019-09533-Z/TABLES/1

Schneider, A., Friedl, M.A., Potere, D., 2009. A new map of global urban extent from MODIS satellite data. Environmental Research Letters 4, 044003. https://doi.org/10.1088/1748-9326/4/4/044003

Schneider, M., Schelte, T., Schmitz, F., Körner, M., 2023. EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union. Scientific Data 2023 10:1 10, 1–10. https://doi.org/10.1038/s41597-023-02517-0

Schroeder, W., Oliva, P., Giglio, L., Csiszar, I.A., 2014. The New VIIRS 375 m active fire detection data product: Algorithm description and initial assessment. Remote Sensing of Environment 143, 85–96. https://doi.org/10.1016/J.RSE.2013.12.008

See, L., Bayas, J.C.L., Lesiv, M., Schepaschenko, D., Danylo, O., McCallum, I., Dürauer, M., Georgieva, I., Domian, D., Fraisl, D., Hager, G., Karanam, S., Moorthy, I., Sturn, T., Subash, A., Fritz, S., 2022. Lessons learned in developing reference data sets with the contribution of citizens: the Geo-Wiki experience. Environmental Research Letters 17, 065003. https://doi.org/10.1088/1748-9326/AC6AD7

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.Y., Milèinski, G., Nikšieč, M., Painho, M., Podör, A., Olteanu-Raimond, A.M.R., Rutzinger, M., 2016. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS International Journal of Geo-Information 2016, Vol. 5, Page 55 5, 55. https://doi.org/10.3390/IJGI5050055

See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A., Perger, C., Schill, C., Zhao, Y., Maus, V., Siraj, M.A., Albrecht, F., Cipriani, A., Vakolyuk, M., Garcia, A., Rabia, A.H., Singha, K., Marcarini, A.A., Kattenborn, T., Hazarika, R., Schepaschenko, M., van der Velde, M., Kraxner, F., Obersteiner, M., 2015. Building a hybrid land cover map with crowdsourcing and geographically weighted regression. ISPRS Journal of Photogrammetry and Remote Sensing 103, 48–56. https://doi.org/10.1016/J.ISPRSJPRS.2014.06.016

Sheppard, S.R.J., Cizek, P., 2009. The ethics of Google Earth: Crossing thresholds from spatial data to landscape visualisation. Journal of Environmental Management 90, 2102–2117. https://doi.org/10.1016/J.JENVMAN.2007.09.012

Shetty, S., Gupta, P.K., Belgiu, M., Srivastav, S.K., 2021. Assessing the Effect of Training Sampling Design on the Performance of Machine Learning Classifiers for Land Cover Mapping Using Multi-Temporal Remote Sensing Data and Google Earth Engine. Remote Sensing 2021, Vol. 13, Page 1433 13, 1433. https://doi.org/10.3390/RS13081433

Shimabukuro, Y., Duarte, V., Anderson, L., Valeriano, D., Arai, E., Freitas, R., Rudorff, B.F., Moreira, M., 2007. Near real time detection of deforestation in the Brazilian Amazon using MODIS imagery. Ambiente e Agua-An Interdisciplinary Journal of Applied Science 1, 37–47. https://doi.org/doi:10.4136/ambi-agua.4

Skidmore, A., 2017. Environmental Modelling with GIS and Remote Sensing. Environmental Modelling with GIS and Remote Sensing. https://doi.org/10.4324/9780203302217
Song, X.P., 2023. The future of global land change monitoring. International Journal of Digital Earth 16, 2279–2300. https://doi.org/10.1080/17538947.2023.2224586
Song, X.P., Hansen, M.C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S.V., Di Bella, C.M., Conde, M.C., Copati, E.J., Fernandes, L.B., Hernandez-Serna, A., Jantz, S.M., Pickens, A.H., Turubanova, S., Tyukavina, A., 2021. Massive soybean expansion in South America since 2000 and implications for conservation. Nature Sustainability 4, 784–792. https://doi.org/10.1038/s41893-021-00729-z
Song, X.P., Hansen, M.C., Stehman, S.V., Potapov, P.V., Tyukavina, A., Vermote, E.F., Townshend, J.R., 2018. Global land change from 1982 to 2016. Nature 560, 639–643. https://doi.org/10.1038/s41586-018-0411-9
Song, X.P., Potapov, P.V., Krylov, A., King, L.A., Di Bella, C.M., Hudson, A., Khan, A., Adusei, B., Stehman, S.V., Hansen, M.C., 2017. National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. Remote Sensing of Environment 190, 383–395. https://doi.org/10.1016/j.rse.2017.01.008
Staquet, M., Rozencweig, M., Lee, Y.J., Muggia, F.M., 1981. Methodology for the assessment of new dichotomous diagnostic tests. Journal of Chronic Diseases 34, 599–610. https://doi.org/10.1016/0021-9681(81)90059-X
Steele, B.M., Chris Winne, J., Redmond, R.L., 1998. Estimation and Mapping of Misclassification Probabilities for Thematic Land Cover Maps. Remote Sensing of Environment 66, 192–202. https://doi.org/10.1016/S0034-4257(98)00061-3
Steele, B.M., Patterson, D.A., Redmond, R.L., 2003. Toward estimation of map accuracy without a probability test sample. Environmental and Ecological Statistics 10, 333–356. https://doi.org/10.1023/A:1025111108050
Stehman, S.V., 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. International Journal of Remote Sensing 35, 4923–4939. https://doi.org/10.1080/01431161.2014.930207
Stehman, S.V., 2013. Estimating area from an accuracy assessment error matrix. Remote Sensing of Environment 132, 202–211. https://doi.org/10.1016/j.rse.2013.01.016
Stehman, S.V., 2012. Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. Remote Sensing Letters 3, 111–120. https://doi.org/10.1080/01431161.2010.541950
Stehman, S.V., 2009a. Sampling designs for accuracy assessment of land cover. International Journal of Remote Sensing 30, 5243–5272. https://doi.org/10.1080/01431160903131000
Stehman, S.V., 2009b. Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. Remote Sensing of Environment 113, 2455–2462. https://doi.org/10.1016/j.rse.2009.07.006
Stehman, S.V., 2004. A Critical Evaluation of the Normalized Error Matrix in Map Accuracy Assessment. Photogrammetric Engineering and Remote Sensing 70, 743–751. https://doi.org/10.14358/PERS.70.6.743
Stehman, S.V., 2001. Statistical Rigor and Practical Utility in Thematic Map Accuracy Assessment. Photogrammetric Engineering & Remote Sensing 67, 727–734.
Stehman, S.V., 2000. Practical implications of design-based sampling inference for thematic map accuracy assessment. Remote Sensing of Environment 72, 35–45. https://doi.org/10.1016/S0034-4257(99)00090-5
Stehman, S.V., 1999. Comparing thematic maps based on map value. International Journal of Remote Sensing 20, 2347–2366. https://doi.org/10.1080/014311699212065

Stehman, S.V., Czaplewski, R.L., 1998. Design and Analysis for Thematic Map Accuracy Assessment - an application of satellite imagery. Remote Sensing of Environment 64, 331–344. https://doi.org/10.1016/S0034-4257(98)00010-8
Stehman, S.V., Fonte, C.C., Foody, G.M., See, L., 2018. Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover. Remote Sensing of Environment 212, 47–59. https://doi.org/10.1016/J.RSE.2018.04.014
Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment 231. https://doi.org/10.1016/j.rse.2019.05.018
Stehman, S.V., Mousoupetros, J., McRoberts, R.E., Næsset, E., Pengra, B.W., Xing, D., Horton, J.A., 2022. Incorporating interpreter variability into estimation of the total variance of land cover area estimates under simple random sampling. Remote Sensing of Environment 269. https://doi.org/10.1016/j.rse.2021.112806
Stehman, S.V., Olofsson, P., Woodcock, C.E., Herold, M., Friedl, M.A., 2012. A global land-cover validation data set, II: augmenting a stratified sampling design to estimate accuracy by region and land-cover class. International Journal of Remote Sensing 33, 6975–6993. https://doi.org/10.1080/01431161.2012.695092
Stehman, S.V., Pengra, B.W., Horton, J.A., Wellington, D.F., 2021. Validation of the U.S. Geological Survey's Land Change Monitoring, Assessment and Projection (LCMAP) Collection 1.0 annual land cover products 1985–2017. Remote Sensing of Environment 265, 112646. https://doi.org/10.1016/J.RSE.2021.112646
Stehman, S.V., Wagner, J.E., 2024. Choosing a sample size allocation to strata based on trade-offs in precision when estimating accuracy and area of a rare class from a stratified sample. Remote Sensing of Environment 300, 113881. https://doi.org/10.1016/j.rse.2023.113881
Stehman, S.V., Wickham, J.D., 2011. Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. Remote Sensing of Environment 115, 3044–3055. https://doi.org/10.1016/J.RSE.2011.06.007
Stehman, S.V., Wickham, J.D., Wade, T.G., Smith, J.H., 2008. Designing a Multi-Objective, Multi-Support Accuracy Assessment of the 2001 National Land Cover Data (NLCD 2001) of the Conterminous United States. photogramm eng remote sensing 74, 1561–1571. https://doi.org/10.14358/PERS.74.12.1561
Stehman, S.V., Xing, D., 2022. Confidence intervals for proportion of area estimated from a stratified random sample. Remote Sensing of Environment 280, 113193. https://doi.org/10.1016/J.RSE.2022.113193
Stöcker, C., Bennett, R., Nex, F., Gerke, M., Zevenbergen, J., 2017. Review of the Current State of UAV Regulations. Remote Sensing 2017, Vol. 9, Page 459 9, 459. https://doi.org/10.3390/RS9050459
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., 2006. Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps. GOFC-GOLD Report No. 25. Office for Official Publications of the European Communities, Luxemburg.
Sulla-Menashe, D., Gray, J.M., Abercrombie, S.P., Friedl, M.A., 2019. Hierarchical mapping of annual global land cover 2001 to present: The MODIS Collection 6 Land Cover product. Remote Sensing of Environment 222, 183–194. https://doi.org/10.1016/J.RSE.2018.12.013
Sullivan, B., 2021. NICFI's satellite imagery of the global tropics now available in Earth Engine for analysis. Medium.
Szantoi, Z., Geller, G.N., Tsendbazar, N.E., See, L., Griffiths, P., Fritz, S., Gong, P., Herold, M., Mora, B., Obregón, A., 2020. Addressing the need for improved land cover map

products for policy support. Environmental Science & Policy 112, 28–35. https://doi.org/10.1016/J.ENVSCI.2020.04.005
Tang, X., Bullock, E.L., Olofsson, P., Estel, S., Woodcock, C.E., 2019. Near real-time monitoring of tropical forest disturbance: New algorithms and assessment framework. Remote Sensing of Environment 224, 202–218. https://doi.org/10.1016/J.RSE.2019.02.003
Tarko, A., Tsendbazar, N.E., de Bruin, S., Bregt, A.K., 2021. Producing consistent visually interpreted land cover reference data: learning from feedback. International Journal of Digital Earth 14, 52–70. https://doi.org/10.1080/17538947.2020.1729878
Thenkabail, P.S., Teluguntla, P.G., Xiong, J., Oliphant, A., Congalton, R.G., Ozdogan, M., Gumma, M.K., Tilton, J.C., Giri, C., Milesi, C., Phalke, A., Massey, R., Yadav, K., Sankey, T., Zhong, Y., Aneece, I., Foley, D., 2021. Global Cropland-Extent Product at 30-m Resolution (GCEP30) Derived from Landsat Satellite Time-Series Data for the Year 2015 Using Multiple Machine-Learning Algorithms on Google Earth Engine Cloud: U.S. Geological Survey Professional Paper 1868. https://doi.org/10.3133/pp1868
Thompson, I.D., Maher, S.C., Rouillard, D.P., Fryxell, J.M., Baker, J.A., 2007. Accuracy of forest inventory mapping: Some implications for boreal forest management. Forest Ecology and Management 252, 208–221. https://doi.org/10.1016/J.FORECO.2007.06.033
Townshend, J., Justice, C., Li, W., Gurney, C., McManus, J., 1991. Global land cover classification by remote sensing: present capabilities and future possibilities. Remote Sensing of Environment 35, 243–255. https://doi.org/10.1016/0034-4257(91)90016-Y
Townshend, J.R., 1992. Land cover. International Journal of Remote Sensing 13, 1319–1328. https://doi.org/10.1080/01431169208904193
Trodd, N.M., 1995. Uncertainty in land cover mapping for modelling land cover change. Proceedings of RSS95 remote sensing in action 1138–1145.
Tsendbazar, N.E., De Bruin, S., Fritz, S., Herold, M., See, L., Roy, P.S., Thenkabail, P.S., 2015a. Spatial Accuracy Assessment and Integration of Global Land Cover Datasets. Remote Sensing 2015, Vol. 7, Pages 15804-15821 7, 15804–15821. https://doi.org/10.3390/RS71215804
Tsendbazar, N.E., de Bruin, S., Herold, M., 2015b. Assessing global land cover reference datasets for different user communities. ISPRS Journal of Photogrammetry and Remote Sensing 103, 93–114. https://doi.org/10.1016/J.ISPRSJPRS.2014.02.008
Tsendbazar, N.E., de Bruin, S., Mora, B., Schouten, L., Herold, M., 2016. Comparative assessment of thematic accuracy of GLC maps for specific applications using existing reference data. International Journal of Applied Earth Observation and Geoinformation 44, 124–135. https://doi.org/10.1016/J.JAG.2015.08.009
Tsendbazar, N.E., Herold, M., de Bruin, S., Lesiv, M., Fritz, S., Van De Kerchove, R., Buchhorn, M., Duerauer, M., Szantoi, Z., Pekel, J.F., 2018. Developing and applying a multi-purpose land cover validation dataset for Africa. Remote Sensing of Environment 219, 298–309. https://doi.org/10.1016/J.RSE.2018.10.025
Tsendbazar, N.E., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., Van De Kerchove, R., Duerauer, M., 2021. Towards operational validation of annual global land cover maps. Remote Sensing of Environment 266. https://doi.org/10.1016/j.rse.2021.112686
Tsendbazar, N.E., Tarko, A., Linlin, L., Herold, M., Lesiv, M., Fritz, S., Maus. V, 2020. Copernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015-2019: Validation Report. Zenodo, Geneve, Switzerland. https://doi.org/10.5281/zenodo.3938974
Tsutsumida, N., Comber, A.J., 2015. Measures of spatio-temporal accuracy for time series land cover data. International Journal of Applied Earth Observation and Geoinformation 41, 46–55. https://doi.org/10.1016/J.JAG.2015.04.018

Turubanova, S., Potapov, P., Hansen, M.C., Li, X., Tyukavina, A., Pickens, A.H., Hernandez-Serna, A., Arranz, A.P., Guerra-Hernandez, J., Senf, C., Häme, T., Valbuena, R., Eklundh, L., Brovkina, O., Navrátilová, B., Novotný, J., Harris, N., Stolle, F., 2023. Tree canopy extent and height change in Europe, 2001–2021, quantified using Landsat data archive. Remote Sensing of Environment 298, 113797. https://doi.org/10.1016/J.RSE.2023.113797

Turubanova, S., Potapov, P.V., Tyukavina, A., Hansen, M.C., 2018. Ongoing primary forest loss in Brazil, Democratic Republic of the Congo, and Indonesia. Environ. Res. Lett. 13, 074028. https://doi.org/10.1088/1748-9326/aacd1c

Tyukavina, A., Baccini, A., Hansen, M.C., Potapov, P.V., Stehman, S.V., Houghton, R.A., Krylov, A.M., Turubanova, S., Goetz, S.J., 2015. Aboveground carbon loss in natural and managed tropical forests from 2000 to 2012. Environmental Research Letters 10, 74002. https://doi.org/10.1088/1748-9326/10/7/074002

Tyukavina, A., Hansen, M.C., Potapov, P., Parker, D., Okpa, C., Stehman, S.V., Kommareddy, I., Turubanova, S., 2018. Congo Basin forest loss dominated by increasing smallholder clearing.

Tyukavina, A., Hansen, M.C., Potapov, P.V., Stehman, S.V., Smith-Rodriguez, K., Okpa, C., Aguilar, R., 2017. Types and rates of forest disturbance in Brazilian Legal Amazon, 2000–2013. Science Advances 3, 1–16. https://doi.org/10.1126/sciadv.1601047

Tyukavina, A., Potapov, P., Hansen, M.C., Pickens, A.H., Stehman, S.V., Turubanova, S., Parker, D., Zalles, V., Lima, A., Kommareddy, I., Song, X.-P., Wang, L., Harris, N., 2022. Global Trends of Forest Loss Due to Fire From 2001 to 2019. Frontiers in Remote Sensing 3. https://doi.org/10.3389/frsen.2022.825190

Tyukavina, A., Stehman, S.V., Pickens, A.H., Potapov, P., Hansen, M.C., 2025. Practical global sampling methods for estimating area and map accuracy of land cover and change. Remote Sensing of Environment 324, 114714. https://doi.org/10.1016/j.rse.2025.114714

Tyukavina, A., Stehman, S.V., Potapov, P.V., Turubanova, S.A., Baccini, A., Goetz, S.J., Laporte, N.T., Houghton, R.A., Hansen, M.C., 2013. National-scale estimation of gross forest aboveground carbon loss: A case study of the Democratic Republic of the Congo. Environmental Research Letters 8.

Valle, D., Izbicki, R., Leite, R.V., 2023. Quantifying uncertainty in land-use land-cover classification using conformal statistics. Remote Sensing of Environment 295, 113682. https://doi.org/10.1016/J.RSE.2023.113682

Vancutsem, C., Achard, F., Pekel, J.F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragão, L.E.O.C., Nasi, R., 2021. Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. Science Advances 7. https://doi.org/10.1126/SCIADV.ABE1603/SUPPL_FILE/ABE1603_SM.PDF

Vancutsem, C., Marinho, E., Kayitakire, F., See, L., Fritz, S., 2012. Harmonizing and Combining Existing Land Cover/Land Use Datasets for Cropland Area Monitoring at the African Continental Scale. Remote Sensing 2013, Vol. 5, Pages 19-41 5, 19–41. https://doi.org/10.3390/RS5010019

Venter, Z.S., Barton, D.N., Chakraborty, T., Simensen, T., Singh, G., 2022. Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover. Remote Sensing 14, 4101. https://doi.org/10.3390/RS14164101/S1

Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sensing of Environment 114, 106–115. https://doi.org/10.1016/J.RSE.2009.08.014

Verhegghen, A., d'Andrimont, R., Waldner, F., van der Velde, M., 2021. Accuracy Assessment of the First Eu-Wide Crop Type Map with Lucas Data. International Geoscience and Remote Sensing Symposium (IGARSS) 2021-July, 1990–1993. https://doi.org/10.1109/IGARSS47720.2021.9553758

Wagner, J.E., Stehman, S.V., 2015. Optimizing sample size allocation to strata for estimating area and map accuracy. Remote Sensing of Environment 168, 126–133. https://doi.org/10.1016/J.RSE.2015.06.027

Waldner, F., Schucknecht, A., Lesiv, M., Gallego, J., See, L., Pérez-Hoyos, A., d'Andrimont, R., de Maet, T., Bayas, J.C.L., Fritz, S., Leo, O., Kerdiles, H., Díez, M., Van Tricht, K., Gilliams, S., Shelestov, A., Lavreniuk, M., Simões, M., Ferraz, R., Bellón, B., Bégué, A., Hazeu, G., Stonacek, V., Kolomaznik, J., Misurec, J., Verón, S.R., de Abelleyra, D., Plotnikov, D., Mingyong, L., Singha, M., Patil, P., Zhang, M., Defourny, P., 2019. Conflation of expert and crowd reference data to validate global binary thematic maps. Remote Sensing of Environment 221, 235–246. https://doi.org/10.1016/J.RSE.2018.10.039

Walker, D.A., Reynolds, M.K., Daniëls, F.J.A., Einarsson, E., Elvebakk, A., Gould, W.A., Katenin, A.E., Kholod, S.S., Markon, C.J., Melnikov, Evgeny S., Moskalenko, Natalia G., Talbot, S.S., Yurtsev, B.A., Bliss, L.C., Edlund, S.A., Zoltai, S.C., Wilhelm, M., Bay, C., Gudjónsson, G., Moskalenko, N. G., Ananjeva, G.V., Drozdov, D.S., Konchenko, L.A., Korostelev, Y.V., Melnikov, E. S., Ponomareva, O.E., Matveyeva, N.V., Safranova, I.N., Shelkunova, R., Polezhaev, A.N., Johansen, B.E., Maier, H.A., Murray, D.F., Fleming, M.D., Trahan, N.G., Charron, T.M., Lauritzen, S.M., Vairin, B.A., 2005. The Circumpolar Arctic vegetation map. Journal of Vegetation Science 16, 267–282. https://doi.org/10.1111/J.1654-1103.2005.TB02365.X

Whitcraft, A., Becker-Reshef, I., Justice, C., Gifford, L., Kavvada, A., Jarvis, I., 2019. No pixel left behind: Toward integrating Earth Observations for agriculture into the United Nations Sustainable Development Goals framework. Remote Sensing of Environment 235.

White, J.C., Coops, N.C., Wulder, M.A., Vastaranta, M., Hilker, T., Tompalski, P., 2016. Remote Sensing Technologies for Enhancing Forest Inventories: A Review. Canadian Journal of Remote Sensing 42, 619–641. https://doi.org/10.1080/07038992.2016.1207484

Witjes, M., Herold, M., de Bruin, S., 2024. Iterative mapping of probabilities: A data fusion framework for generating accurate land cover maps that match area statistics. International Journal of Applied Earth Observation and Geoinformation 131, 103932. https://doi.org/10.1016/j.jag.2024.103932

Witjes, M., Parente, L., van Diemen, C.J., Hengl, T., Landa, M., Brodský, L., Halounova, L., Križan, J., Antonić, L., Ilie, C.M., Craciunescu, V., Kilibarda, M., Antonijević, O., Glušica, L., 2022. A spatiotemporal ensemble machine learning framework for generating land use/land cover time-series maps for Europe (2000-2019) based on LUCAS, CORINE and GLAD Landsat. PeerJ 10, e13573. https://doi.org/10.7717/PEERJ.13573/FIG-19

WMO, 2022. The 2022 GCOS ECVs Requirements. World Meteorological Organization, Geneva, Switzerland.

Woodcock, C.E., Gopal, S., 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. International Journal of Geographical Information Science 14, 153–172. https://doi.org/10.1080/136588100240895

Woodcock, C.E., Loveland, T.R., Herold, M., Bauer, M.E., 2020. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. Remote Sensing of Environment 238, 111558. https://doi.org/10.1016/J.RSE.2019.111558

Woodcock, C.E., Ozdogan, M., 2012. Trends in Land Cover Mapping and Monitoring, in: Gutman, G., Janetos, A.C., Justice, C.O., Moran, E.F., Mustard, J.F., Rindfuss, R.R., Skole, D., Turner, B.L., Cochrane, M.A. (Eds.), Land Change Science. Springer, Dordrecht, pp. 367–377. https://doi.org/10.1007/978-1-4020-2562-4_21

Wu, B., Zhang, M., Zeng, H., Tian, F., Potgieter, A.B., Qin, X., Yan, N., Chang, S., Zhao, Y., Dong, Q., Boken, V., Plotnikov, D., Guo, H., Wu, F., Zhao, H., Deronde, B., Tits, L., Loupian, E., 2023. Challenges and opportunities in remote sensing-based crop

monitoring: a review. National Science Review 10. https://doi.org/10.1093/NSR/NWAC290
Wulder, M.A., Coops, N.C., Roy, D.P., White, J.C., Hermosilla, T., 2018. Land cover 2.0. International Journal of Remote Sensing 39, 4254–4284. https://doi.org/10.1080/01431161.2018.1452075
Wulder, M.A., White, J.C., Goward, S.N., Masek, J.G., Irons, J.R., Herold, M., Cohen, W.B., Loveland, T.R., Woodcock, C.E., 2008. Landsat continuity: Issues and opportunities for land cover monitoring. Remote Sensing of Environment 112, 955–969. https://doi.org/10.1016/j.rse.2007.07.004
Xiao, X., Dorovskoy, P., Biradar, C., Bridge, E., 2011. A library of georeferenced photos from the field. Eos, Transactions American Geophysical Union 92, 453–454. https://doi.org/10.1029/2011EO490002
Xing, D., Stehman, S.V., 2024. Using interpenetrating subsampling to incorporate interpreter variability into estimation of the total variance of land cover area estimates. Remote Sensing of Environment 311, 114289. https://doi.org/10.1016/J.RSE.2024.114289
Xu, G., Zhu, X., Fu, D., Dong, J., Xiao, X., 2017. Automatic land cover classification of geo-tagged field photos by deep learning. Environmental Modelling & Software 91, 127–134. https://doi.org/10.1016/J.ENVSOFT.2017.02.004
Xu, P., Herold, M., Tsendbazar, N.E., Clevers, J.G.P.W., 2020. Towards a comprehensive and consistent global aquatic land cover characterization framework addressing multiple user needs. Remote Sensing of Environment 250, 112034. https://doi.org/10.1016/J.RSE.2020.112034
Xu, P., Tsendbazar, N.E., Herold, M., de Bruin, S., Koopmans, M., Birch, T., Carter, S., Fritz, S., Lesiv, M., Mazur, E., Pickens, A., Potapov, P., Stolle, F., Tyukavina, A., Van De Kerchove, R., Zanaga, D., 2024. Comparative validation of recent 10 m-resolution global land cover maps. Remote Sensing of Environment 311, 114316. https://doi.org/10.1016/J.RSE.2024.114316
Yao, H., Qin, R., Chen, X., 2019. Unmanned Aerial Vehicle for Remote Sensing Applications—A Review. Remote Sensing 2019, Vol. 11, Page 1443 11, 1443. https://doi.org/10.3390/RS11121443
Ying, Q., Hansen, M.C., Potapov, P.V., Tyukavina, A., Wang, L., Stehman, S.V., Moore, R., Hancher, M., 2017. Global bare ground gain from 2000 to 2012 using Landsat imagery. Remote Sensing of Environment 194, 161–176. https://doi.org/10.1016/j.rse.2017.03.022
Zalles, V., Hansen, M.C., Potapov, P.V., Parker, D., Stehman, S.V., Pickens, A.H., Parente, L.L., Ferreira, L.G., Song, X.P., Hernandez-Serna, A., Kommareddy, I., 2021. Rapid expansion of human impact on natural land in South America since 1985. Science Advances 7. https://doi.org/10.1126/SCIADV.ABG1620/SUPPL_FILE/ABG1620_SM.PDF
Zalles, V., Harris, N., Stolle, F., Hansen, M.C., 2024. Forest definitions require a re-think. Commun Earth Environ 5, 620. https://doi.org/10.1038/s43247-024-01779-9
Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.E., Xu, P., Ramoino, P., Arino, O., 2022. ESA WorldCover 10 m 2021 v200 [Dataset]. https://doi.org/10.5281/zenodo.7254221
Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., Tsendbazar, N.-E., Ramoino, F., Arino, O., 2021. ESA WorldCover 10 m 2020 v100 [Dataset]. https://doi.org/10.5281/zenodo.5571936

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., Mi, J., 2021. GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery. Earth System Science Data 13, 2753–2776. https://doi.org/10.5194/ESSD-13-2753-2021

Zhu, Z., Woodcock, C.E., 2014. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. Remote Sensing of Environment 152, 217–234. https://doi.org/10.1016/J.RSE.2014.06.012

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., Pekel, J.F., Lymburner, L., Pahlevan, N., Scambos, T.A., 2019. Benefits of the free and open Landsat data policy. Remote Sensing of Environment 224, 382–385. https://doi.org/10.1016/J.RSE.2019.02.016

Zimmerman, P.L., Housman, I.W., Perry, C.H., Chastain, R.A., Webb, J.B., Finco, M.V., 2013. An accuracy assessment of forest disturbance mapping in the western Great Lakes. Remote Sensing of Environment 128, 176–185. https://doi.org/10.1016/j.rse.2012.09.017