Ref. Ares(2024)8516585 - 29/11/2024

Evoland
LAND MONITORING EVOLUTION

D3.8 Prototype Validation Plan - final version

Work Package 3 – Prototype Demonstration

Authors: Justine Hugé, Alexandre Pennec (CLS)

Date: 30.11.2024

Funded by the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Full Title	**Evolution of the Copernicus Land Service portfolio** (integrating novel EO data and latest machine learning algorithms to continuously monitor the status, dynamics and biomass of the land surface)		
Project number	101082130	**Acronym**	EVOLAND
Start date	01.01.2023	**Duration**	36 months
Granting authority	European Health and Digital Executive Agency (HaDEA)		
Project Coordinator	Vlaamse Instelling voor Technologisch Onderzoek N.V. (VITO)		
Date of delivery	**Contractual** 21	**Actual**	23
Type	R - Document, report	**Dissemination level**	SEN - Sensitive
Lead beneficiary	CLS		
Lead author	Justine Hugé, Alexandre Pennec	**Email**	jhuge@groupcls.com apennec@groupcls.com
Other authors	Linda Moser (GAF AG), Andreas Langner (GAF AG), Heinz Gallaun (JR), Janik Deutscher (JR), Martin Puhm (JR), Wanda De Keersmaecker (VITO), Kasper Bonte (VITO), Astrid Vannoppen (VITO), Nick Gutkin (VITO)		
Reviewer(s)	Tim Ng (VITO), Ruben Van De Kerchove (VITO)		
Keywords	Validation plan, Prototype demonstration activities, Thematic proof-of-concept		

Document Revision History				
Version	**Issue date**	**Stage**	**Changes**	**Contributor**
0.1	29.10.2024	Draft	First interim version	CLS, GAF, JR, VITO
1.0	30.11.2024	Final	Review and finalization	CLS

Disclaimer
Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Copyright message
© EvoLand consortium, 2024
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgment of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Acronyms

Acronym	Description
AD	Applicable Document
AGB	Above Ground Biomass
AI	Artificial Intelligence
ALS	Airborne laser scanning
AOI	Area Of Interest
BFAST	Break Detection for Satellite Time Series Data
CAP	Common Agricultural Policy
CCI	Climate Change Initiative
CDSE	Copernicus Data Space Ecosystem
CGLOPS	The Global Land Component of the Copernicus Land Service
CGLS	Copernicus Global Land Service
CI	Confidence Interval
CLC	Corine Land Cover
CLMS	Copernicus Land Monitoring Service
CwRS	Control with Remote Sensing
DEM	Digital Elevation Model
DLT	Dominant Leaf Type
DMP	Dry Matter Production
DSM	Digital Surface Model
EC	Eddy Covariance
ECV	Essential Climate Variable
EEA	European Environment Agency
EFFIS	European Forest Fire Information System
EO	Earth Observation
EU	European Union
ESA	European Space Agency
ETC	European Topic Centres
EVOLAND	Evolution of Copernicus Land Monitoring Service
EWMA	Exponentially Weighted Moving Average
FAO	Food and Agriculture Organization
FCH	Forest Canopy Height
GCOS	Global Climate Observing System
GLAD	Global Land Analysis and Discovery
GEDI	Global Ecosystem Dynamics
GEE	Google Earth Engine
GFW	Global Forest Watch
GPP	Gross Primary Production
GSWE	Global Surface Water Extent
GSWO	Global Surface Water Occurrence
HRL	High Resolution Layer
IACS	Integrated Administration and Control System
ICOS	Integrated Carbon Observation System
IMD	Imperviousness Degree
INSPIRE	Infrastructure for Spatial Information in Europe

LACO	Land Cover Validation Platform
LAEA	Lambert Azimuthal Equal-Area projection
LC	Land Cover
LIDAR	LIght Detection And Ranging
LPIS	Land Parcel Information System
LSC	Land surface categories
LU	Land Use
LUE	Light Use Efficiency
LUCAS	Land Use/Cover Area frame statistical Survey
LULC	Land Use and Land Cover
MAE	Mean Absolute Error
MARS	Monitoring Agriculture with Remote Sensing
MEA	Maximum Eligible Area
MMU	Minimum Mapping Unit
MMW	Minimum Mapping Width
NA	Not Applicable
NDVI	Normalized Difference Vegetation Index
NFI	National Forest Inventory
NPP	Net Primary Production
NRT	Near Real Time
NVLCC	Non Vegetated Land Cover Characteristics
OA	Overall Accuracy
OBIA	Object-Based Image Analysis
OSM	OpenStreetMap
PA	Producer Accuracy
PSU	Primary Sampling Unit
QA	Quality Assessment
QC	Quality Control
REDD	Reduction
RMSE	Root Mean Square Error
SAR	Synthetic Aperture Radar
SLF	Small Landscape Feature
SSU	Secondary Sampling Unit
SWF	Small Woody Feature
SWOT	Surface Water Ocean Topography
TBC	To Be Confirmed
TBD	To Be Determined
TCD	Tree Cover Density
TCDM	Tree Cover Disturbance Monitoring
TPCP	Tree Cover Presence Change
UA	User Accuracy
USA	United States of America
VHR	Very High Resolution
VLCC	Vegetated Land Cover Characteristics
WP	Work Package

Table of Contents

Table of Tables

Table of Figures

Applicable Documents

ID	Applicable document
[AD1]	Grant Agreement
[AD2]	D1.1 Service requirements report - interim version
[AD3]	D1.3 EO and in-situ data requirements report - interim version
[AD4]	D3.1 EvoLand prototypes report - interim version
[AD5]	D3.7 Prototype Validation Plan – interim version
[AD6]	D2.8 Existing and novel in-situ and training datasets - interim version 2

Executive Summary

The Prototype demonstration activities (WP3) for the EvoLand (Evolution of Copernicus Land Monitoring Service) project build upon methods developed in WP2, serving as thematic proof-of-concept for advancements in the Copernicus Land Monitoring products. Eleven next-generation CLMS services are pre-selected within the EvoLand project. WP3 aims to produce prototypes for specific thematic topics across 8 test sites, predominantly in the EEA-38+UK territory, covering diverse bio-geographic regions. Each prototype will be tested in at least three sites to assess suitability and performance, addressing topics such as Forest, Agriculture, Water, Urban, and General Land Cover. Feasibility, calibration measures, and upscaling will be analysed and tested. All prototypes will undergo statistically sound accuracy assessments handled independently within Task 3.6. The goal is to demonstrate potential feasibility, accuracy, and usability improvements for future Copernicus Land Monitoring products.

This document is the final version (D3.8) of the Prototype Validation Plan for the EvoLand project. It aims to describe the specific validation approach (sampling design, adequate metrics, etc.) implemented for each prototype and provides updates to its previous interim version (D3.7).

The document lays the groundwork for the comprehensive testing that will unfold in the coming months as part of Phase 2, particularly with the advent of product availability and continuous monitoring. The validation protocol for the prototypes is outlined and was refined in the current document according to the outcomes of the Phase 1. It has to be noted that as Phase 2 is now started, it is possible that the final product descriptions will evolve and will be established as Phase 2 ends, with for a number of prototypes, the description of the classes, the input and validation data set, the target quality and the detailed validation process evolving.

This final validation plans due for November 2024, encompassing test methods, lessons learned and recommendations for each prototypes in terms of validation procedures, will serve as the foundation for a comprehensive and operational service.

1 Introduction

1.1 About EvoLand

The Evolution of the Copernicus land service portfolio integrating novel EO data and latest machine learning algorithms to continuously monitor the status, dynamics and biomass of land surface (EvoLand) project (Proj. nr: 101082130) develops innovative methods, algorithms and candidate Copernicus Land Monitoring Service (CLMS) prototypes monitoring the status and changes of land use/land cover and various land surface characteristics at high spatial and temporal resolution. Under the new Copernicus 2.0 programme, CLMS will continue to provide products and services for the 2021-2027 period, thus building time series that enable for long term change monitoring. Additionally, harmonisation and further integration of the products across components is envisaged as well as a more efficient workflow by further automation, modularity of the software and data products, and the increased use of Artificial Intelligence to provide real and near real time data processing, to respond to emerging European policy needs.

In short EvoLand will:

- provide tangible proof of the evolution potential of the CLMS in terms of improved information content, quality and timeliness;
- produce up to eleven operationally benchmarked and qualified candidate CLMS monitoring the status and changes of land use/land cover and various land surface characteristics at high resolution;
- develop new innovative methods and algorithms (data fusion, continuous monitoring, AI, biomass mapping, novel infrastructure, EO and in-situ data); and
- analyse policy, data and infrastructure requirements for the prototype services.

1.2 About work package 3: Prototype demonstration

The Prototype demonstration activities [WP3] is built on the methods developed under WP2. They will be the thematic proof-of-concept of these developments demonstrating their potential feasibility, accuracy and potential usability and improvements within the future Copernicus Land Monitoring products. Eleven candidates next generation CLMS services have been pre-selected within the EvoLand project (see Table 1).

WP3 foresees the production of prototypes for specific Thematic topics in eight test sites (see section 3.2.3) mainly in the EEA-38+UK territory (six in continental Europe and one in French Guiana) plus one in Africa (Côte d'Ivoire), covering various bio-geographic regions and landscapes in a representative manner. Each prototype, defined in the sub-sections below, will be tested / implemented in at least three sites. The WP3 aims at assessing the suitability and performance of this setup for generating several prototype products with an expected accuracy and validity.

In WP3, the feasibility of the products, the identification of potential calibration measures and a later upscaling of these products will also be analysed and tested. It will address the topics below, each one covered by different prototypes and sub-topics: Forest (Task 3.1), Agriculture (Task 3.2), Water (Task 3.3), Urban (Task 3.4) and General Land Cover (Task 3.5). All prototypes will be subject to statistically sound accuracy assessments/estimations following a dedicated internal validation by the consortium, which will be handled in an independent way within the frame of the Task 3.6.

1.3 About this document

This document is the **final version** (D3.8) of the **Prototype Validation Plan** for the EvoLand project.

It aims to describe the specific validation approach (sampling design, adequate metrics, etc.) implemented for each prototype (Table 1) developed in WP3.

Table 1. Candidate next-generation Copernicus Land Monitoring Services

Theme	Prototype title
Forest	C1: Continuous forest monitoring
	C2: Forest disturbance mapping
	C3: Forest biomass mapping
Agriculture	C4: Cover crop type mapping
	C5: Cropland/grassland GPP mapping
	C6: Small landscape features mapping
Water	C7: Improved water bodies extent
Urban	C8: Continuous imperviousness monitoring
	C9: Automated Land Use mapping of urban dynamics
General land cover	C10: Continuous mapping of land surface categories
	C11: On-demand land cover mapping

The validation plan can be subdivided into three main steps:

i. Stratification and sampling design
ii. Response design (visual interpretation of reference samples on VHR data (QA/QC breakpoint) or in-situ data/ancillary data collection; and
iii. Statistical analysis, accuracy assessments (Producer's Accuracy, User's Accuracy, Overall Accuracy, Kappa, F-Score) and reporting at the various required levels.

The design of validation approach will follow state-of-the-art best practices on validation but will be tailored to each prototype. An independent validation will be performed for each of them with reference data not used in the production process.

The document is structured as follows:

- Section 1 (this section) introduces the document and the EvoLand project.
- Section 2 describes the general concepts of the validation process, from generalities on state-of-the-art best practices, sampling design, response design and statistical analysis.
- Section 3 investigates the validation plan for each prototype.
- Section 4 gives a conclusion.
- Section 5 reports the references of publications used in the current document.

The previous document (D3.7) presented the general concepts of the validation and provided a first insight into how each prototype will be validated based on the latest scientific validation and statistical standards, the status of the project and the specifications of the prototypes.

The current final version of this document (D3.8) is an update of D3.7 and provides full details on the validation plan with regard to the updated specifications of the prototypes. At the end of Phase 1, some prototypes already show a correct level of maturity while for some of them, methodologies still need to be improved during Phase 2 with potentially description to be updated at the same time. In this regard, the validation plan that is described in this report may need to be slightly adjusted as Phase 2 ends.

This report introduces for all prototypes, a section related to the lesson learnt, taking advantages on the experience gained during the Phase 1 in terms of validation. A section related to the recommendations for operationalization is also included, which brings some insight into a possible operationalization of the given prototype.

2 Validation: general concepts

In the context of the project, a list of prototypes has been derived from the user requirements to support the evolution of the Copernicus Land Monitoring Service to better meet key user needs and policy objectives. Section 3.1 presents in detail this list of prototypes.

The prototypes can be traditionally divided into 2 categories:

- The **status** prototypes describe the current state or condition of the land cover/use, the surface physical characteristic, provide information about existing land cover types and their distribution at a specific point in time but doesn't capture changes over the time,
- The **change** prototypes focus on detecting changes over the time period and so offer insights into trends like urbanization, deforestation disturbances.

The status and change prototypes include:

- The **continuous** products: products with continuous thematic class values such as tree cover density, imperviousness degree, above ground woody biomass, forest canopy height,
- The **Boolean** products: products with discrete thematic class values such as tree/non-tree cover, cover crop type, water bodies extent.

However, in the context of the project, the consortium will develop new methods to move from the traditional status and change focused land cover maps towards continuous monitoring of land cover dynamics. The continuous monitoring will include time series, Near Real-Time (NRT) products:

- **Time series** involves the analysis of land cover/use over a continuous period at different time intervals to create a temporal sequence.
- **Near Real-Time** (NRT) products provide information that is close to real-time, often with minimal delay. The NRT products aim to capture the most recent changes in land cover/use.

Each prototype with its own specifications will be validated accordingly the best practices described in the next sections.

2.1 Generalities or best practices

This validation plan is built according to the latest scientific validation and statistical standards, so that the validation approach is fully compatible with the state-of-the-art peer-reviewed literature in the field of validation. Stehman & Foody (2019) outline six good practice criteria for accuracy assessment:

1. **Map relevant criterion** – the accuracy assessment should address the reality of the landscape mapped, i.e., the accuracy estimates, and confusion matrix should reflect the proportional areal representation of the study area.
2. **Statistically rigorous** – assuming design-based inference, a probability sampling design should be used, and variability of the accuracy and area estimates should be quantified by reporting standard errors or confidence intervals.
3. **Quality assured** – reference data and results should be monitored and evaluated.
4. **Reliable** – a repetition of the accuracy assessment protocol, with a different sample and different interpreters, should produce approximately similar results. It is important to report standard errors or confidence intervals. Accuracy and area estimates should have small standard errors, and there should be small variability among interpreters when assigning reference class labels.
5. **Transparent** – provide all relevant details to readers to inform on the quality of the results.
6. **Reproducible** – the methodology should be reproducible given the documentation provided.

The overall methodology for assessment of the thematic accuracies will follow these principles to ensure high quality procedures. Our validation framework is presented in Figure 1 detailing the different steps which will be carried out and that are presented in the following sections.

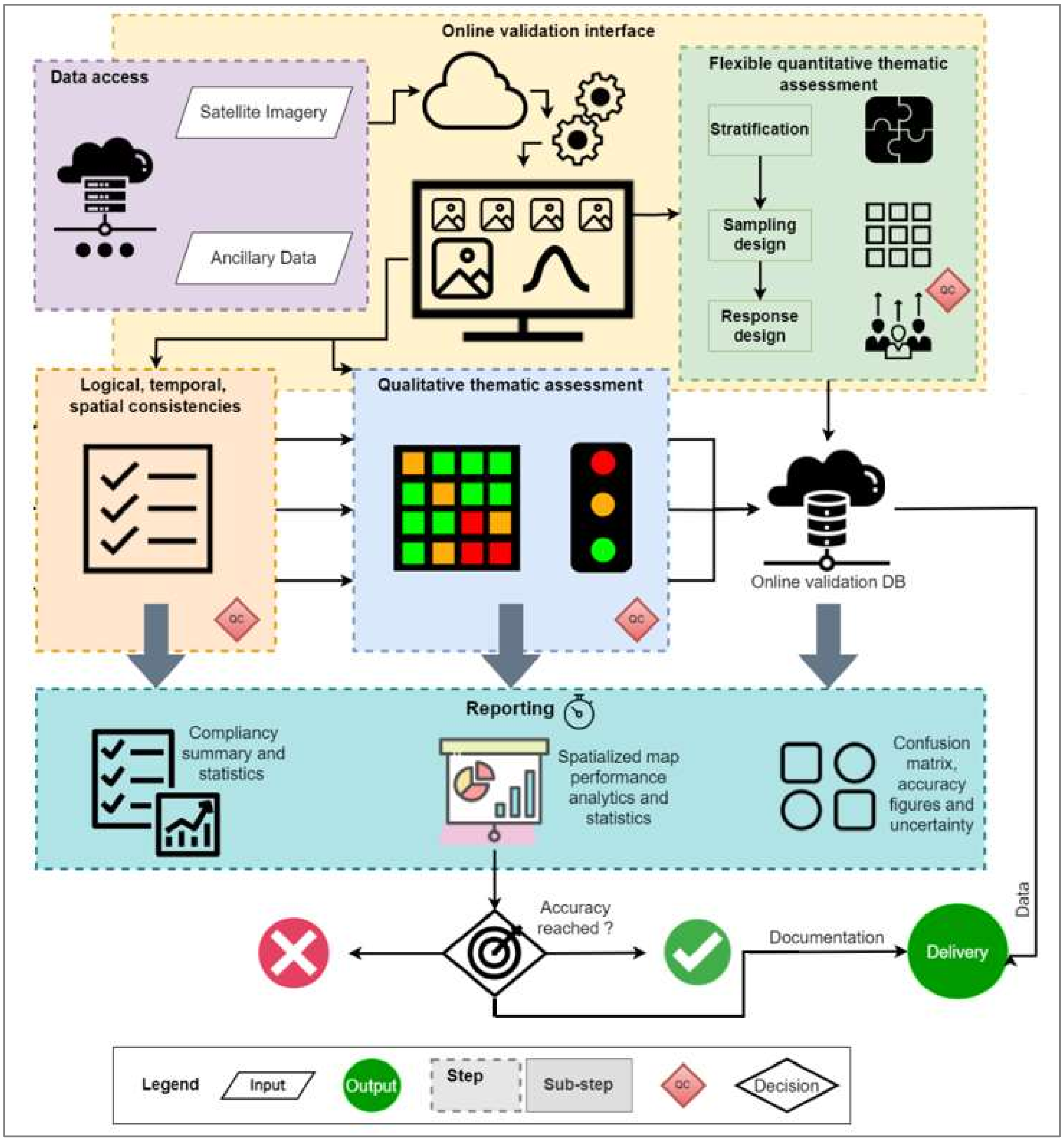


Figure 1. General validation framework.

All prototypes will be validated using a transparent traceable validation procedure relying on statistical quantities, **independent** from the production process to be considered as scientifically credible. The proposed validation procedure follows protocols approved by Infrastructure for Spatial Information in Europe (INSPIRE) directives. The independence of the validation process adopted in the proposed approach is two-fold:

- Reference datasets used for the validation will not be used as part of the production of the products to be validated. The accuracy assessment is conducted independently.

- When possible, the validation will be carried out by teams not involved in the production of the Prototypes.

The overall methodology for assessment of the thematic accuracies is fully explained in the following chapters.

2.2 Qualitative accuracy assessment

A **qualitative thematic accuracy** procedure can be integrated in the validation process consisting of quality control of the products. Although of good overall quality, the geospatial outputs may exhibit macroscopic errors such as tile and scene effects which are not acceptable, sharp features alterations due to omission and commission errors (including but not limited to land cover borders, and other physical elements such as riverbeds, coastline, bare soil and rocks especially in mountain areas etc.).

Accordingly, the validation plan recommends a qualitative thematic accuracy assessment protocol in addition to a quantitative accuracy assessment (see, for example Szantoi et al., 2021).

The products can be examined at sampled cell-level of a regular grid. The pattern delineation and artefacts can be assessed and compared to Very High Resolution (VHR)/High Resolution (HR) imagery and other existing EEA and Copernicus products to highlight macroscopic errors (commission and omission).

Results can be stored in a database associating each cell to an identification code related to a predefined typology of errors and the errors coordinates (see for example Figure 2). The type, number and gravity of errors per cell are summarized into a categorical legend ranging from:

- Correct cells (in green): no qualitative thematic issues were identified,
- Minor errors (in yellow): one or two thematic issues were identified in the cell,
- Reasonable errors (in orange): two or three thematic issues were identified, and
- Major errors (in red): three or five thematic issues were identified.

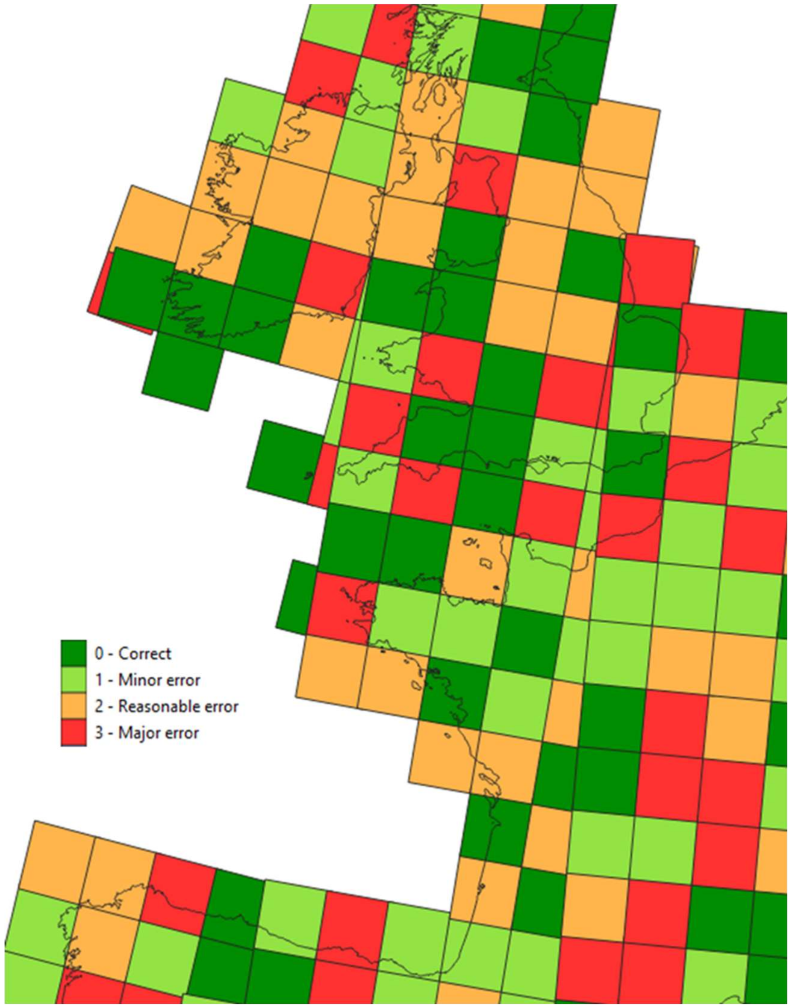


Figure 2. Example of georeferenced database for the qualitative assessment.

However, the qualitative thematic accuracy assessment cannot be used to provide thematic accuracy metrics and there is a need for a quantitative thematic accuracy assessment of the prototypes.

2.3 Quantitative accuracy assessment

2.3.1 Sampling design and stratification

2.3.1.1Sampling design

Sampling design refers to the protocol whereby the samples are selected.

Within the validation process of each prototype, accuracy assessment objectives include precise estimates by class. A **probability sampling design** is essential for map validation (Olofsson et al., 2013; Stehman et al., 2009), to provide unbiased metrics, and will be applied, when possible, in the present project. However, the protocol will be adapted to the specifications of the prototypes and a probability sampling may not be possible to be implemented for each product mainly due to the availability of reference data. Section 3 will present the validation plan for each individual prototype.

A probabilistic sampling is one in which the probability of inclusion of a sample can be determined. Typically, **simple random, stratified random, clustered random and systematic designs** are all examples of probability sampling designs (as shown in Figure 3 and Figure 4). It can also be achieved through a random or systematic approach in combination with a stratification, as shown by the early implementation of satellite-based crop area estimation systems in the USA (Allen, 1990) and in Europe as part of the Monitoring Agriculture with Remote Sensing (MARS) project in the late 1980s and early 1990s (Taylor, 1997).

In simple random designs, classes covering a small portion of the population may not be adequately sampled. Clustered sampling is often used to reduce the costs of the collection of reference data but does not resolve geographic distribution problems. Stratified approaches overcome this drawback and apply prior knowledge of the study area to first divide the area into strata. Therefore, the stratified random sampling of points is one of the most common approaches to assess map thematic accuracy but may not be the best option for assessing areas. Stratified systematic with random origin has the advantage of enhancing traceability and is a better solution for assessing areas. The main limitation of systematic sampling is that there are no unbiased estimators of the variance, and the simple random sampling variance is often used, which slightly overestimates the true value (Bellhouse, 1988).

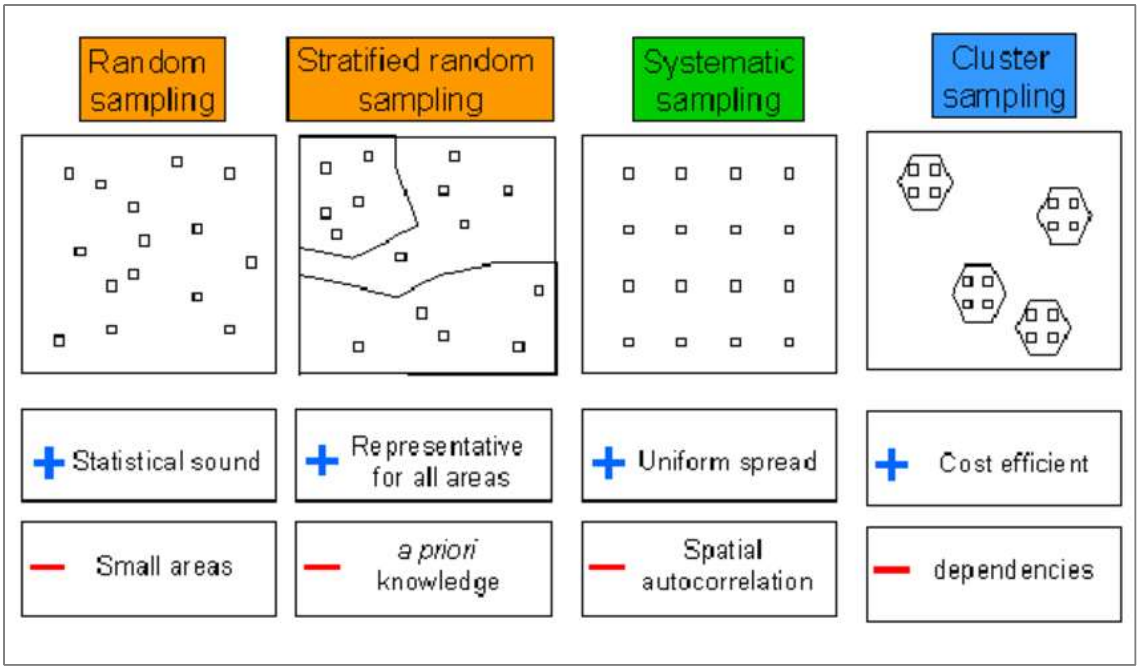


Figure 3. Different types of sampling design with advantages and disadvantages (source: G. Banko, 1998)

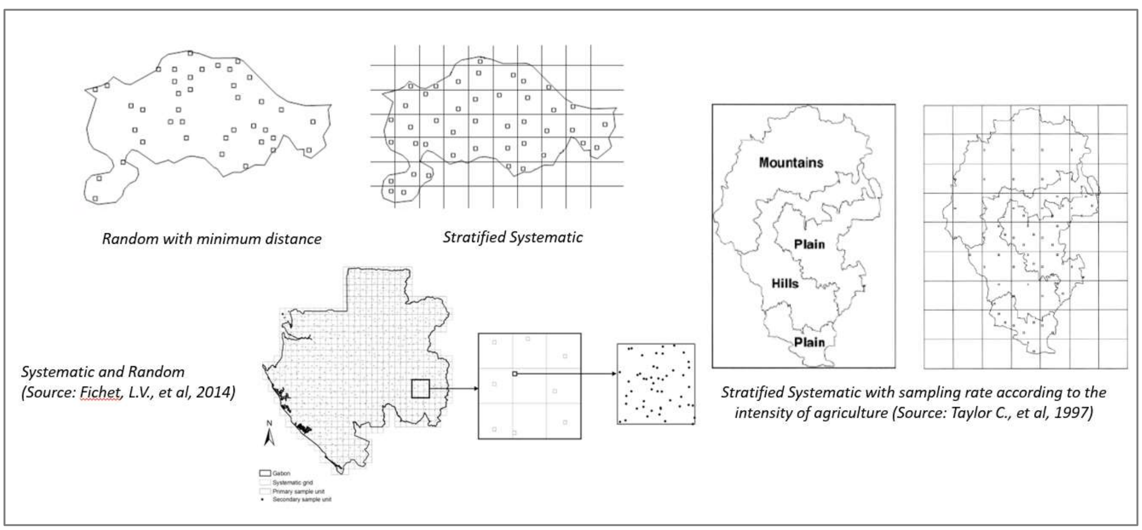


Figure 4. Different types of combined sampling design.

In summary, a probabilistic sample approach is needed for unbiased estimation and will be employed in the present project when feasible. **Following guidelines from (Olofsson et al., 2014; Stehman & Foody, 2019), the consortium has thus selected a stratified random sampling design which combines benefits from both protocols**.

2.3.1.2 Stratification

Stratification is the partition of a study area into several non-overlapping distinct subsets called strata. Stratification is useful in this context of validation because:

- It allows a better representation of all the thematic classes,
- it reduces the amount of effort required for the reference data collection, and
- it optimizes the sample distribution to increase precision of estimates.

In many cases, the land cover map to be assessed or validated is directly used as stratification (Lowell and Jaton, 2000). This is a good approach to estimate commission errors in a binary classification but may be weak for other cases. If we have some information that quantifies the likely proportion of errors, for example a measure of landscape complexity, it can be a more efficient stratification for all types of errors. We shall choose a higher sampling rate in more difficult areas, where both the errors and their variances are higher.

Stratification is also possible using external datasets combined as basis for stratification including a combination of physical information (e.g., Digital Elevation Model (DEM)), existing land cover maps and agro-climatic conditions (Agro-Ecological Zones) for example.

In any case, experience has shown that a too complex stratification will not bring major improvement and that there must be a clear case for stratifying the Area of Interest (AOI). A good stratification should reduce the variability within each stratum, and considering the complexity of patterns, which may be difficult to accurately describe. A too complex and insufficiently accurate stratification may be counterproductive and lead to substantial variability with a wider 95% confidence interval for example of the resulting crop area estimates as compared with simple random sampling. In cases where agro-climatic conditions are difficult to identify based on available data, production statistics per administrative unit can be used. However, statistics at some administrative level should be used carefully as described by Sannier et al. (1998).

Statistics and biophysical factors can be combined in a relatively complex analysis to stratify the area, but experience in Europe from the Regional Inventory programme of the MARS project in the 1990s (Taylor, 1997) showed that increasing complexity for analysis to stratify the area, increases the inefficiency of the stratification. In addition, the total number of strata should not be large because, even if the stratification allows to improve the efficiency of the sampling by reducing the number of sample units per stratum, the total number of sampling units may be larger if too many strata are present. Pragmatically, Cochran (1977, p. 134) recommends no more than 6-8 strata in total.

In conclusion, considering good practices from the literature and past experience, we recommend selecting no more than 6 to 8 strata when applying a stratified sampling.

2.3.2 Response design

The response design specifies aspects on the collection of the reference data considered as surrogate to "ground truth" and subsequently interpreted by experts (or directly collected in the field during a survey). The aim is to collect data that represent the real condition on the ground, which can then be compared with the classified map to derive accuracy statistics (Congalton & Green, 2009).

Key aspects involve the selection of (i) the sample unit: pixel, square cluster of pixels or polygons (Radoux and Bogaert, 2017), (ii) source of reference data and how it is collected and interpreted, and (iii) how agreement between the map and reference classification is defined (e.g., product legend and MMU). These choices strongly influence the results of the accuracy assessment (Stehman & Foody, 2019; Stehman & Wickham, 2011).

(i) Sample size per stratum

A critical feature of stratified sampling designs is the sample size and its sample allocation to strata, as different allocations favour different estimation objectives (Stehman & Foody, 2019). Sample size calculations are based on the objectives of the assessment and the desired precision requirements. **The number of samples selected is a trade-off between the number needed for statistical validity, and the cost of collecting reference data** (Congalton & Green, 2009; Stehman & Foody, 2019).

When using stratified sampling, the main goal is to maximise the efficiency of the stratification by optimising the sample allocation per strata with a view to (1) minimise the overall number of samples necessary (by reducing the number of samples for strata) and/or (2) minimise the uncertainty of the resulting estimates.

In the case of stratified sampling, samples must be assigned to a stratum once the overall sample size has been determined. **The allocation of samples only affects precision (i.e., standard error) of the estimators, not the bias** (Olofsson et al., 2014). Equal, proportional, and optimal allocations exist, each favouring different estimation objectives (Stehman & Foody, 2019). A simple way is the use of equal allocation, but this is usually not very efficient. Proportional allocation (to the fraction of the total area covered by a given strata) is an option, but it will give disappointing results for classes considered important but covering a small proportion, e.g., crop types covering a small area present only in very specific strata.

It is possible to **estimate a suitable sample size for each stratum based on the expected acceptable error rate** which can be referred to the error rate requested by the technical specifications and or previous test results decision (Olofsson et al., 2014; Stehman & Foody, 2019). The standard error of the error rate can be calculated as follows:

$$\sigma_h = \sqrt{\frac{p_h(1-p_h)}{n_h}}$$

where n_h is the sample size for stratum h and p_h is the expected error rate.

This can be reworked to express the sample size n_h as a function of p_h and desired standard error σ_h:

$$n_h = \frac{p_h(1-p_h)}{\sigma_h^2}$$

From Figure 5, it can be seen that for an expected 50% error rate, within a stratum, 100 samples would be required to guarantee a standard error of 5%, whereas the number of samples would need to be increased by a factor of four if the accepted standard deviation is divided by a factor of 2. On the other hand, if the expected error rate is 15%, only 51 samples would be necessary with a 5% standard error. A similar approach was adopted to determine the sample size for assessing the accuracy of CORINE Land Cover (CLC) 2006 and CLC 2000-2006 changes (Büttner et al., 2012).

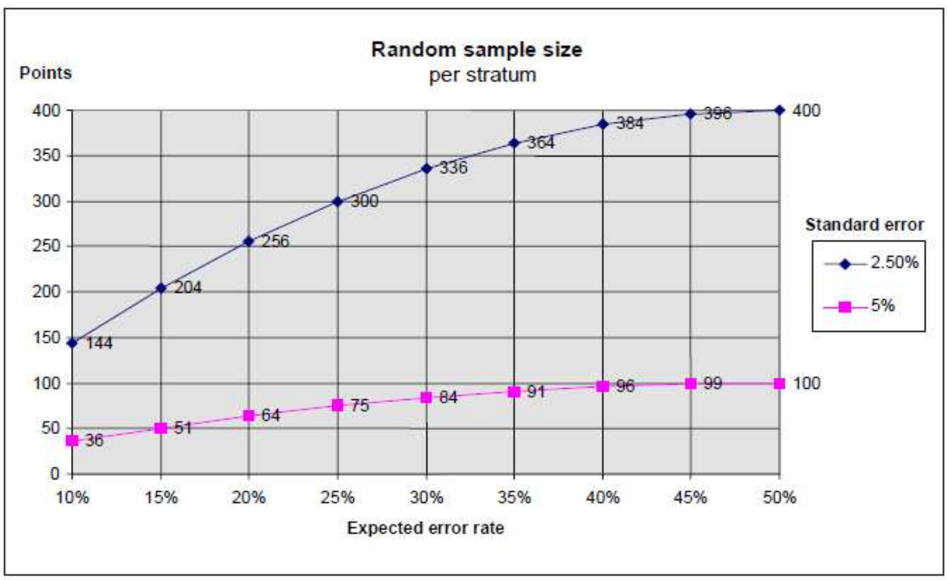


Figure 5. Number of sample points as a function of the expected error rate for two accepted standard error values (after Wack et al., 2012).

The efficiency of the sample allocation per stratum can be optimized. Stehman (2009) advises using the Neyman optimal allocation to estimate the area of change and the overall accuracy (Cochran, 1977):

$$n_h = n * (N_h * \sigma_h) / [\Sigma(N_i * \sigma_i)]$$

where n_h is the sample size for stratum h, n is the total sample size, N_h is the population size for stratum h, σ_h is the standard error of the error rate of stratum h, I extends to all strata within the AOI.

According to Stehman (2012), Neyman optimal allocation should be preferred for estimating area as well as overall accuracy, whereas equal allocation is effective for estimating user's accuracy.

In practice (Särndal et al., 1992, pp. 267 and 407) recommend a minimum within-stratum sample sizes of 10-20; (Cochran, 1977, p. 134) recommends minimum within-stratum sample sizes of 20; and for temperate forest inventories (Westfall et al., 2011) recommend within-stratum sample sizes of at least 20.

(ii) reference data

Reference data for validation should be of higher quality than the data used for creating the map. If using the same source, the process of creating the reference classification should be more accurate than the process of creating the map (Congalton & Green, 2009; Olofsson et al., 2014; Stehman & Foody, 2019; Stehman & Wickham, 2011). The process of creating the reference classification should also be independent of the map in order to avoid a biased estimate of the accuracy (Radoux and Bogaert, 2020).

There are several options to obtain high quality reference dataset. One option relates to the collection of existing datasets. The reference will come from the best datasets available for the given regions to be analysed. Another one refers to the processing and photointerpretation of Very High spatial Resolution (VHR) images. Indeed, high-spatial resolution aerial photography, and Object-Based Image Analysis (OBIA) techniques, provide new methods to accurately and rapidly extract information at a low cost (Zhang et al., 2018).

It is known that for a large part of the globe, already existing datasets of good quality are available. Some are available at local scale, some are available at National scale or at global level. An exhaustive list of additional potential reference datasets is available in the interim deliverable for the **Existing and novel in-situ and training datasets** [AD3, AD6]. Some are official data from authoritative public bodies (thus, supposed to be exhaustive and consistent). In close relation with the production team, a special attention will be given to the data set that will be used for calibration. **The validation team will in no way use the same dataset**. When possible, the consortium proposes to use VHR images to control the reference data sets selected and to complement it if required. The advantages in doing so are:

- better fit the reference data to the period of production,
- be independent of any training set that could be used in the production phase,
- have a human control of the data used as a reference.
- assure that the reference dataset is not incomplete, and
- assure that the reference data set comes from data of relevant spatial resolution.

A **double-blind approach** is recommended:

- the validation is undertaken without considering the map layer information, and
- the production is done without knowledge of the final interpreted validation data.

A **plausibility analysis** can then be implemented considering the information of the map layer. It will contribute to provide an in-depth understanding of the causes of classification errors and to account for ambiguous map values at the class borders. During the plausibility analysis, the sample Units in disagreement are re-interpreted (i.e., Product vs. First blind interpretation) and the expert evaluates whether the classification of samples is plausible or not. The interpretation is carried out knowing the blind interpretation and the map products classification.

To conduct the interpretation, a dedicated and customized internal software interface can be used for image interpretation and image management. The consortium also considers using already existing tools available online. Some of them have been even used supporting internal projects, such as:

- LACO-Wiki LandCover validation platform from IIASA.
- GeoWiki
- Collect Earth with access to Google Earth Engine
- Google Earth Engine

The tool will be chosen at the discretion of the partner responsible for the interpretation.

In the case where an existing database is used as reference data to estimate the quality of the product, it may be useful to cite the work of Xu (Xu et al., 2024). The authors compared 5 methods for validating global land cover products, which allow the quality of the spatial representation and thematic quality to be estimated, considering the uncertainties associated with the reference data. The spatial detail of the maps was assessed at different levels of homogeneity within a 3 × 3 kernel. As indicated in the paper, of the approaches considered, validation using the sample label supplemented by the majority label within the neighborhood is found to produce more reasonable accuracy estimates compared to the overly optimistic approach of using any label within the neighborhood and the overly pessimistic approach of direct comparison between map and reference labels. **Such methods could be used in the future operationalization of some of the prototypes.** It will be checked if such an approach can be used for the prototyping phase.

2.3.3 Statistical analysis

The last step in the production of the mapping consists in analysing the samples in order to draw conclusions about the accuracy of the products.

Continuous products

Concerning the continuous prototypes i.e., product with continuous thematic classes values (tree cover density, imperviousness degree, above ground woody biomass, forest canopy height), the following accuracy metrics can be calculated:

- **The Root Mean Square Error (RMSE)** is a quadratic scoring rule that also measures the average magnitude of the error. The RMSE is calculated based on the square root of the average of squared differences between reference and map observation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \bar{\mathrm{y}}_\mathrm{i})^2}$$

- **The Mean Absolute Error (MAE)** measures the absolute average magnitude of the errors. The MAE is calculated based on the average over the controlled sample dataset of the absolute differences between reference data and classified observation where all individual differences have equal weight.

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} . \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{y_i} - \bar{\mathrm{y}}_\mathrm{i}|$$

- A scatterplot of the fraction share values extracted from the sample units for both the reference and map data will be made with a view to assess the **correlation between reference and map values** and identify any systematic bias (slope and intercept of the regression line significantly different for 1 and 0 respectively).

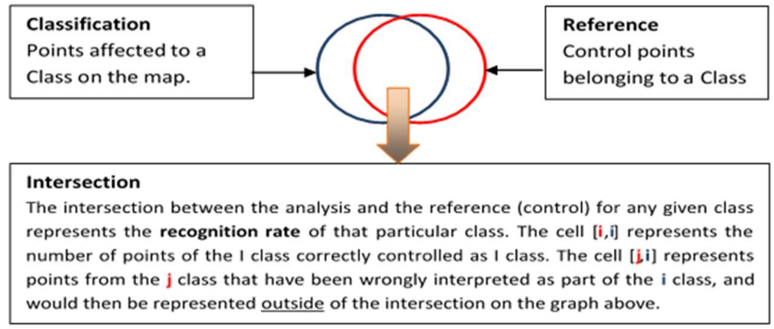
Boolean products

For the Boolean prototypes i.e., products with discrete thematic classes values, thematic accuracy is presented in the form of an error matrix made from the results of the sample units (in-situ data or visually interpreted). As explained in Selkowitz and Stehman (2011), the **sampling intensity** (defined as the proportion of the population that is been sampled, not to be confused with **sampling size** that refers to the number of sample units) resulting from a stratified systematic sampling approach is not equal for all strata. These differences should be accounted for by applying a weight factor to each sample unit within a given stratum. Therefore, a correction for the sampling intensity will be applied to the error matrices produced following the procedure described by Selkowitz and Stehman (2011) and applied by Olofsson et al. (2013) leading to a weighting factor inversely proportional to the inclusion probability of samples from a given stratum. Not applying this correction could result in underestimating or overestimating map accuracies. The final sampling intensities is calculated for each stratum based on the ratio between the number of samples included in the validation dataset and the size/total area of the stratum considered and the weight factor (*p*) will be calculated according to the following equation:

$$\hat{p}_{ij} = \left(\frac{1}{N}\right) \sum_{x \in (i,j)} \frac{1}{\pi_h^*}$$

Where *i* and *j* are the columns and rows in the error matrix, *N* is the total number of possible units (population) and π_h^* is the sampling intensity for a given stratum.

Then, the construction of the error matrix is described as illustrated in Figure 6 for five thematic classes.



	Class i	Class j	Class k	Class l	Class m	Total Classification
Class i	i i	i j	i k	i l	i m	Total Ci
Class j	j i	j j	j k	j l	j m	Total Cj
Class k	k i	k j	k k	k l	k m	Total Ck
Class l	l i	l j	l k	l l	l m	Total Cl
Class m	m i	m j	m k	m l	m m	Total Cm
Total Reference	Total (Ri)	Total (Rj)	Total (Rk)	Total (Rl)	Total (Rm)	Overall Total (T)

Figure 6. Confusion matrix for accuracy assessment of thematic map products.

Referring to Figure 6, let us assume α represents any given class of [*I, j, k, l, m*], the following accuracy metrics can be calculated:

- The **Overall Accuracy** or **Recognition Rate** is measured by the sum of the diagonal of the Confusion Matrix divided by the total number of controlled points: $OA\ or\ R_r = \sum_{\alpha=i}^{m}(\alpha\alpha)/\mathrm{T}$. The **Recognition Rate** or **Overall Accuracy** assesses the overall agreement between the classified and reference data set. However, for single class themes, it does not necessarily provide a realistic assessment of the quality of the map produced because there can be substantial unbalance between omission and commission errors.
- Therefore, the row and column totals and the diagonal of the Matrix are used to assess two types of accuracy, the User's and Producer's Accuracy:
  - **Producer Accuracy** for the α class = $\alpha\alpha/C\alpha$ is a measure of **omission error**. For instance, an observation has been identified as Tree Cover in the validation set but has been classified as another class: it has been omitted from the forest class.
  - **User Accuracy** for the α class = $\alpha\alpha/R\alpha$ is a measure of the **commission error** (or contamination risk): errors due to the wrong allocation of an observation to a class. For instance, an observation is classified as

forest but identified as belonging to another class during the validation process: this observation has contaminated another class.

- The **F-Score** is a statistical measure that combines user and producer accuracies into a single metric providing a balanced assessment of a classification's performance for a given class.

The standard error of the overall, producer and user accuracies can be calculated as follows:

$$\sigma_h = \sqrt{Var_h} = \sqrt{\frac{p_h(1-p_h)}{n_h}}$$

where n_h is the sample size for stratum h and p_h is the expected error rate.

The total standard error is then the square root of the sum of the variance times the square of the area D of each stratum:

$$\sigma_{Total} = \sqrt{D_h^2 . Var_h}$$

The 95% confidence interval ($CI_{95\%}$) which indicates that the "true " value has a 95% probability to be included within the calculated range, is then defined as:

$$CI_{95\%} = \pm 1.96 * \sigma_{Total}$$

2.3.4 Continuous product validation

For products providing regular information updates (e.g., monthly or annually), the methods discussed in previous sections can be applied to assess data quality. However, additional methods may be explored to enable continuous validation of generated information. EvoLand provides innovative candidates prototypes, and therefore, validation should also provide insights towards innovative processes specially for the continuous products. Users such as the climate modelling community or land managers needs to obtain products that can be inter-compared on a more or less long-term period of time; but the products should also ideally be validated without much time lag after a new release. As mentioned by Tsendbazar et al. (2021) in their study which aims to validate yearly land cover maps during a period of five years, considering that human interpretation is costly and time-consuming, validation datasets should be designed to be easily adjustable for timely and continuous validation of new releases of cover products and yet maintaining statistical rigor. They propose an approach for continuous map validation that includes an accuracy assessment of each map release, evaluating both accuracy stability and temporal consistency. This approach aims to meet user needs for reliable, consistent information on the quality of Global Land Cover maps over time.

It's important to note that applying this method requires access to reference data that matches the desired time frame, geographic coverage, and level of detail. If such data are not readily available, it may be necessary to create reference data through frequent updates using, for example, very high-resolution (VHR) imagery.

Using this method relevant for some prototype within the EvoLand prototyping phases appears unfeasible due to limited availability of reference data across test sites, including those in Europe, and the substantial effort required. Nevertheless, it is included here as an example of an innovative validation protocol that builds upon the core concepts outlined in section 2. This protocol, while challenging to implement at this stage, illustrates a forward-looking approach to validation grounded in established principles.

First consideration is that the first release of the product (T0) must be validated according to the usual method so that the accuracy is well known and because this release T0 will form the basis for next releases of the products (T1, T2…).

Ideally, all releases should be validated with a full exercise, but time and budget may constrain future users from undertaking such an effort.

So, the second consideration is that all the releases of the product (T1, T2…) cannot be validated with the same amount of effort as for T0; and therefore, a hybrid approach will be used.

Tsendbazar et al. (2021) provides insights towards operational validation of annual global land cover maps, which could be adapted to validate EvoLand continuous prototypes when possible. They introduced a multi-purpose validation dataset and building upon this dataset, they developed an operational validation protocol suitable for continuous land cover monitoring at a global scale, herein considering cost-effectiveness and timeliness. They propose a random and targeted reviews of the validation dataset with a collection of extra validation sample sites in possible change areas. The approach described below is partially based on this paper.

Validation of the product release T0

The products are derived from High Resolution (HR) data (e.g., Sentinel-2 at 10m) and mostly needs Very High-Resolution data (VHR) to extract samples for both the calibration of the classification and the validation.

To get a representative reference data set, a two-stage sampling design with Primary Sampling Unit (PSUs) and Secondary Sampling Units (SSUs) is applied. The samples should represent the heterogeneity within the product's thematic e.g., different types of forests that are present in the area and should not omit key areas.

In addition, the sampling design will be used to validate several releases of the products and therefore, should be appropriate to validate the changes in the status of the product for the next releases. Areas of changes are often locally concentrated but should also be represented in the samples. Therefore, a stratified approach is the most adapted sampling. Another point is that the sampling of the area is dependent on the location of the available VHR imagery, if any available.

The PSU size is 1 hectare, and within each PSU 100 SSUs of a size of 10m side (one Sentinel-2 pixel) are located. For each SSU, the interpreter will validate the thematic information e.g., presence of tree cover or non-tree cover, on VHR imagery.

The stratification will be based on the product (release T0).

Validation of T1, T2 releases

Validating the new releases of the product means validating changes between different product versions. In their study, Tsendbazar et al. (2021) propose a partial revisit of the validation dataset consisting of a targeted revisit and a random revisit of the validation dataset as shown in Figure 7.

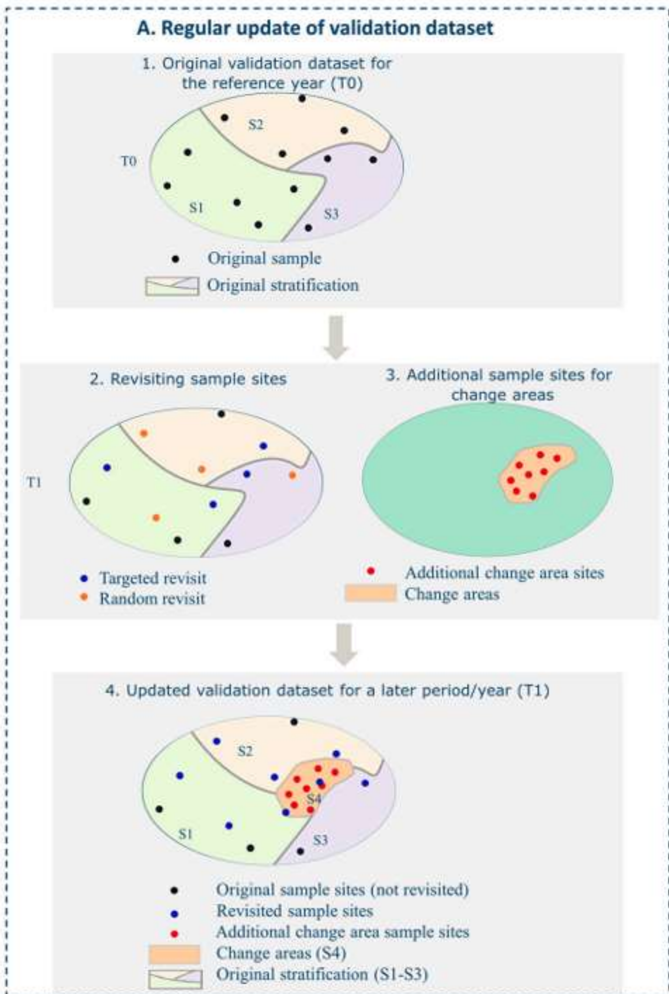


Figure 7. Validation framework envisaged for the continuous products (Source: Tsendbazar et al. (2021)).

We suppose that using the same methodology to derive the different releases, the quality of the stable areas remains the same.

In the case of tree cover / non-tree cover product for example, the changes appearing between the releases can be:

- T0: tree cover → T1: non-tree cover (loss)
- T0: non-tree cover → T1: tree cover (gain)
- The other pixels are being stable between the releases.

It is expected that the most frequent changes will be tree cover to non-tree and adding validation samples into these pixels where classes has been updated means looking for **commission** errors. As demonstrated by Stehman et al. (2012), a stratified random sampling design is suitable for augmenting the sample size. Olofsson et al. (2014) recommend increasing the sample size for rare change classes to achieve an acceptable standard error for estimated user's accuracies and to allocate the remaining sample size roughly proportional to the area occupied by the common classes. Olofsson et al., (2014) recommend using a minimum of 50 samples for the rare strata to reduce standard error of class-specific accuracy estimates.

For the search of **omission** errors, a hybrid approach can be used. Indeed, it is also important to be able to detect any omission in the new release of the product. Increasing the samples size into the rare class as mentioned above, is recommended but is likely to create a situation with one or a few very small strata and one or two very large strata (Olofsson et al. 2020). In addition, the magnitude of the omission error depends on the weight and sample size of the stratum in which the error occurred.

In their study dealing with the assessment of the Near Real Time canopy disturbance product Arevalo et al. (2019) illustrated the **use of a buffer stratum** to overcome this issue, and **to reduce the weight of omission errors** when validating such a product. The authors constructed a buffer stratum around each disturbance pixel in the no-disturbance stratum, allowing to carry less area weight. The hypothesis behind incorporating a spatial buffer into the stratification is that omissions of change typically occur in close proximity to areas of mapped change, while areas mapped as stable forest at larger distances from mapped change are unlikely to contain omissions. Olofsson et al. (2020) have been looking for the best compromise to be reached regarding the size of the buffer, comparing standard errors of deforestation estimates for the different buffer strata and sample sizes in the forest stratum. For a Landsat-based classification with pixel size of 30-meters, the optimal size was found to be seven pixels for 2.000 samples allocated to the forest stratum. The authors also recommend that the optimal buffer size should be at least ten pixels for a large stable class and a small deforestation class. A comparable approach for detailed validation of large-scale Sentinel-2-based forest disturbance maps across Germany was developed by Reinosch et al. (2024).

This buffer approach could easily be applied to the prototype, following the different recommendations, and we would get the following stratification to evaluate the new releases of the prototypes:

- <u>Stratum error of commission</u>: validation dataset is expended with samples randomly selected in the areas with updated classes in the release.

- Stratum high probability of error of omission: samples are randomly selected within a buffer stratum around the identified changes in the product.
- Stratum low probability of error of omission: samples are randomly selected outside the buffer stratum. Another solution could be also to run NDVI index in a time series of Sentinel-2 at the location of the other samples and to detect any breaks in the signal, as performed by Doblas et al. (2020) or to target the revision focusing on validation sample locations that were identified as possibly changed using the BFAST time series algorithm as performed by Tsendbazar et al. (2021).

3 Prototypes Validation Plan

This section provides updated information and validation plan for all EvoLand prototypes. During Phase 1 of the project, many methodological tests were carried out for each prototype to derive the information that led to tailored description. Therefore, the validation plan was not up to date in the previous version of this deliverable [AD5]. This deliverable D3.8 is written as Phase 2 starts. Therefore, while it benefits from the inputs of the different prototypes that occurred during Phase 1, the final and complete description of the validation plan cannot be fully final since updates will be implemented during Phase 2.

3.1 List of Prototypes

In the context of the project, a list of prototypes has been derived from the user requirements, that are candidates for an operationalisation at European scale. Each core thematic prototypes have been tested during Phase 1 and will be updated for Phase 2 and allow to address the different user requirements in terms of thematic classes and temporal resolution. After Phase 1, the list of products has been updated and is provided in Figure 8. Schematic representation of the prototype thematic, responsible partner, and temporal resolution achieved during Phase 1.

, which synthesise also the responsible partners and up-to-date information on the intended temporal frequency target.

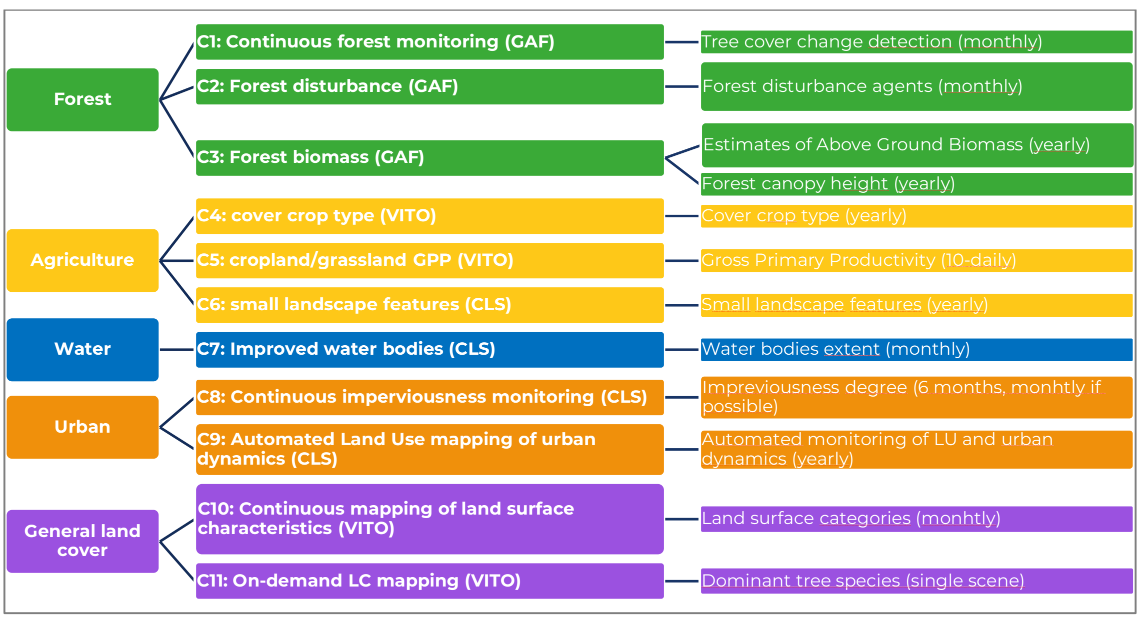


Figure 8. Schematic representation of the prototype thematic, responsible partner, and temporal resolution achieved during Phase 1.

3.2 General concepts

3.2.1 Qualitative accuracy assessment

A qualitative thematic accuracy procedure will be integrated in the validation process. The general concept of the task is presented in section 2.2 and will not be further discussed in the following sections as no adaptations are required.

3.2.2 Quantitative accuracy assessment

Section 2 of this report lists the best practices for handling validation processes and presents the state of the art in this area. However, due to time constraints, available budget and technical background of the partners, **the validation of the prototypes should provide a reasonable trade-off between ideal validation workflow and realistic/pragmatic workflow** that is adapted to the specifications of the prototypes and available in-situ data, while maintaining a scientific approach. The difference between the two workflows can be synthesized as follows (non-exhaustive):

Ideal validation workflow:

- Validation is done by a company that is outside the consortium.
- Complete validation of all the temporal updates of the products.
- A consistent, complete, and timely updated reference dataset is available for all the area or is done manually (response design).

Realistic/pragmatic workflow:

- Identification of validation datasets that can be updated on a regular basis with limited effort to reduce the manual effort needed.
- Validation is done within the consortium as planned in the offer, and a hybrid approach is undertaken with a part of the validation process undertaken by the producer while a part of the validation is done by another group of persons. As mentioned in sections 2.2 and 2.3, the quality of the prototypes products is estimated with quantitative and qualitative accuracy assessments, for which the effort can be shared between partners and this according to the technical expertise. The independence of the validation is still guaranteed with a person not involved in the production that can review of a percentage of the samples used in the response design (quantitative assessment) and a qualitive assessment.

In addition, this deliverable presents the validation plan that will be used for the candidates prototypes as part of the project but also bring insight into the validation that can be used for a further operationalisation of the candidates. While the Consortium has extensive experience in validating pan-European products, the experience gained during the validation of the candidates during the first phase are used to feed the final version of the deliverable, considering lesson learnt. **The Consortium emphasises that for operationalisation of the candidates' prototypes, a validation approach that is closer to the ideal workflow should be used**.

3.2.3 Sites and level of reporting

The prototypes are tested in eight representative demonstration sites mainly in the EEA-38 + UK territory (six in continental Europe and one in French Guiana) plus one in Africa (Côte d'Ivoire), covering various bio-geographic regions and landscapes. **Each prototype will be tested / implemented in at least three sites** (see Figure 9).

The test sites have been optimised to cover in a feasible manner the regional diversity of Land Cover/Land Use (LC/LU), while considering the availability and accessibility of in-situ data.

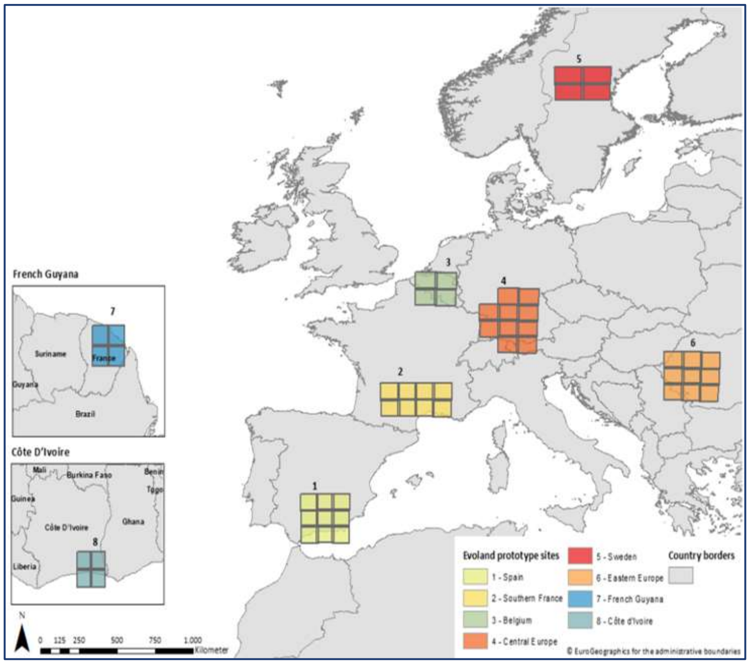


Figure 9. Sites for the validation of the prototypes

Accuracy of the products will be reported at EvoLand prototype site level to assess any potential regional differences based on biogeographical regions or countries for example. Due to the availability of in-situ data for some prototypes, it may be difficult to produce accuracy values for all the demonstration sites. It will be discussed when needed in the current document.

3.2.4 Temporal resolution

Depending on the specific aim of the characteristics to be monitored, the prototypes present different temporal resolutions, namely: continuous, monthly and yearly.

The data for the prototype developments mainly covers data from **2020 – 2022**.

3.2.5 Standardized Prototype Validation Framework

A standardized approach to detailing validation methods, metrics, and results enables to compare products effectively and ensures that each product's validation is transparent and replicable. This consistency not only supports informed decision-making but also establishes a clear benchmark for evaluating performance and improvements across the range of products. In this regard, the **description of the validation that will be done during the Phase 2 will have to report the following level of details for all prototypes**:

<u>General information</u>

- Expected accuracy
- Updated description of the product's thematic classes
- What is the validated information e.g., date of changes, thematic, boundaries of information

Sampling design

- Refer to section 2.1 of the present document for best practices to use,
- Used datasets, satellites images
- Sampling design (if any) applied and rationale behind
- Number of samples (also per strata if applicable) and map representing the geographical location of the samples in the study area

Response design

- How comparison between the product and reference data is defined
- Data used as reference data with temporal information, spatial resolution and how to access to this data.

Analytics

- Refer to section 2.1 of the present document for best practices
- Map displaying the final product with legend and examples of errors and/or improvements

Final analysis

- In terms of methodology to perform validation: explain any deviation from D3.8 guidelines
- Identified challenges in the different sites related to the prototype thematic
- Qualitative analysis
- Quantitative analysis: analysis of results e.g., level of commission, omission

3.3 C1: Continuous Forest Monitoring

3.3.1 Description

The Continuous Forest Monitoring prototype (Table 2) aims to expand and enhance the portfolio and scope of forestry products, particularly the current HRL Vegetated Land Cover Characteristics (VLCC) Forest products. By increasing the update frequency of forest masks, it will capture dynamic changes within tree-covered areas and serve as an input for the forest disturbance agent classifier (C2). This enhanced information supports forest management, planning, forest-related decision-making, and EU policy reporting on forest targets.

Table 2: Technical and thematic specifications for the C1 prototype

Prototype C1	
Prototype	C1 – Continuous Forest Monitoring
Thematic specifications	Time of detection of tree cover canopy disturbance
Prototype sites	Phase 1: Spain (1), Central Europe (4) Phase 2: Spain (1), Central Europe (4), Sweden (5), French Guiana (7)
Temporal frequency	Monthly
Spatial resolution	10m
MMU	Pixel-based
Envisaged accuracy	Will be defined during Phase 2

Currently, following development phase 1, the C1 prototype utilizes the Exponentially Weighted Moving Average (EWMA) method (Brooks et al., 2014), which effectively captures subtle disturbances with small spectral magnitude. However, this sensitivity introduces a trade-off with over-detection, potentially leading to commission errors, which are also connected to insufficient cloud screening which is still to be improved in Phase 2. Notably, early spring months (e.g., March, April) show a tendency for elevated over-detection. The unsupervised EWMA approach does not require training data but does demand extensive Sentinel-2 time series data — spanning a minimum of three full years — before monthly analysis can commence. This requirement involves substantial pre-processing, including cloud masking. Nevertheless, the approach is robust as it is feasible to process large data amounts while still being computationally effective, and in addition it builds up on an open-source tool that has been recommended for forest monitoring applications. On top of the newly introduced monthly updates, compared to the existing 3-yearly forest change product within the VLCC portfolio, the C1 prototype demonstrates promising improvements. This progress suggests significant advantages towards potential yearly change products.

3.3.2 Prototype validation

During the Phase 1, the validated information was the detected changes (change/no change) in the tree cover canopy, and the date of detection of the change.

3.3.2.1 Qualitative assessment

A qualitative accuracy procedure is integrated into the validation process, which consists of product quality control to look for any macroscopic errors. The product provides information on a monthly basis and the quality of detection is first checked in its entirety. As this is a prototyping phase, it takes a reasonable amount of time to check the product at the test sites. The accuracy of the identified change was assessed using Sentinel-2 time-series data, Google Earth imagery and the HLR VLCC layer, the latter providing annual information on tree cover density (TCD). For example, a decrease in the HRL VLCC TCD product was identified to correspond with a disturbance to the tree canopy in the C1 prototype.

Such a qualitative assessment must be carried out during the prototyping phase and a map of the product should be provided in the report.

3.3.2.2 Quantitative assessment

3.3.2.2.1 Sampling design

For validation a stratified systematic sampling design is applied in order to account for different strata sizes while at the same time ensuring full traceability of the results compared to a stratified random sampling approach.

By applying a forest mask (VLCC 2018) in combination with tree cover change information a total of three strata are derived. For that purpose, ancillary data on tree cover change (Tree Cover Loss results 2019-2021 (Hansen et al., 2013), LUCA results 2019-2021 (Mullissa et al., 2021), the upcoming HRL Tree Cover Presence Change (TPCP) 2018-2021[1]) as well as the C1 EWMA results 2019-2022 combined with the TCDM-radar results 2019-2021 (masked with the VLCC 2018 forest mask) are combined and buffered by 100m. The thus derived dataset is intersected with the VLCC 2018 forest mask to split the forest layer into a stratum of stable forest (with no detected changes) and another stratum of forest change:

- Forest change: That stratum allows to estimate the error of commission (producer accuracy). As this stratum includes above mentioned buffer zone it further allows analysing omission errors (producer accuracy) as omissions of change typically occur near areas of mapped change.
- Stable forest (without any detected change): That stratum covers all non-changed areas with a low probability of omission errors, thus contributing to the evaluation of the producer accuracy.
- Stable non-forest: That stratum provides information about the quality of the HRL 2018 forest mask.

As the area of change is comparably small to the unchanged area, this stratified approach allows to reduce the impact of omission errors.

[1] EEA (2025), Copernicus Land Monitoring Service – High Resolution Layer - Tree Cover & Forests, D1.12 HRL Tree Cover & Forests Product User Manual – not yet published

In a preliminary step, with a view to optimise that stratification and sampling allocation process, a systematic grid at regular spaces with about 1000 sampling plots is derived over the whole study area to deliver a first evaluation of changes that occurred during the analysis period. An analysis of the EWMA results provides first information about the expected overall accuracy, which then can be used to calculate the total number of sample points for the following stratified systematic sampling approach following the good practices paper of Olofsson et al 2014. Based on the variance per stratum the optimal allocation of sample points per strata is eventually derived, which follow a stratified systematic sampling design to allow full traceability (systematic sample units are fixed and cannot be moved).

A weighted confusion matrix is used to derive the accuracy values, which allows analysing the results according to the surface representativeness of each stratum.

3.3.2.2.2 Response design

The VHR reference data (VHR_Image_2021, Google Earth time series as well as ESRI World Imagery), for which the acquisition date is known at every pixel location, is visually interpreted by blind analysis considering the product specifications (detection date of the EWMA values). For each sample point, the interpreter can choose between three classes:

- Non-tree: Should be used if the first VHR scene within the analysis period shows non-tree cover – independent of any regrowth at a later stage.
- No change: This is set if the first VHR scene within the analysis period shows tree cover and further VHR scenes with at least one-year different acquisition date do not show any changes.
- Change: This is set if the first VHR scene within the analysis period shows tree cover and further VHR scenes with at least one-year different acquisition date show changes. Changes can be of this nature:
  - Tree to Shrub
  - Tree to No tree
  - Tree to Defoliated tree

After going through all sample points, a confusion matrix is derived showing the raw accuracies. For the sample points where the EWMA result deviates from the reference data, another interpretation undertaken by a second interpreter is carried out as part of a plausibility analysis. Even though above-described blind interpretation approach ensures independence from the disturbance map product, there might be some underestimation of accuracy for difficult points because of shadow effects due to different viewing geometry and sun positions. In a similar way points can be difficult to interpret due to seasonal effects (e.g. leaf shedding) – within deciduous forests. Geolocation issues of the VHR reference scenes as well as geometric inaccuracies of the side-viewing radar data can further cause problems in assigning the correct reference class. This means that the second interpreter should additionally consider the surrounding pixels as well, allowing a one-pixel shift (3x3 kernel) if most of the surrounding pixels show the correct result (majority rule of Xu et al., 2024). Eventually, a consensus must be found with the second interpreter ruling over the first interpreter.

3.3.2.2.3 Analytics

As presented in section 2.3.3, User, Producer, Overall accuracies, 95%CI are calculated, using weighting factors per stratum, which stems from the ratio between the area of the respective stratum and the total area of the study site.

3.3.3 Lesson learned from validation in Phase 1

During Phase 1, when the production of the prototypes started, and in relation to the content of [AD5], it was envisaged to carry out a continuous validation of the monthly updates of the prototypes using a stratified random systematic sampling approach. However, the unavailability of a timely and consistent database made it difficult to follow such an approach. Instead of the continuous validation of the monthly updates, a stratified regular systematic sampling approach was adopted to validate the change map aggregated over the analysis period (e.g. three years for the Central European as well as Spanish test site).

Since this is an entirely new product, in phase 1 some points were manually selected and interpreted in order to get to know the product's performance, hence diverging from a strictly statistical approach as it should be carried out in phase 2. The final product was stratified to create the strata for sampling.

The sampling design was appropriate to study the heterogeneity of the study area. The method still needs to be tested in order to estimate the possible level of quality. In order to calculate the total number of sample points for the final stratified systematic sampling approach it is necessary to try to optimize the change stratum (EWMA and TCDM-radar values in combination with ancillary datasets of Tree Cover Loss and LUCA results). This should be done in such a way that the area of that stratum does not get unnecessary large but at the same time it should cover the maximum number of sample points of the regular grid (preliminary analysis step) that were identified as showing some change during the analysis period.

This can be obtained by testing different (reasonable) sizes of buffer zones around the changed areas – a step on which more emphasis should be put. On the one side an optimal buffer zone will help to derive the correct total number of sample plots, thus allowing a more precise accuracy, but it also eventually reduces the weight of the omission stratum.

3.3.4 Operationalisation: recommendations

In theory, there are two ways to consider the validation of this prototype which provides a continuous update (monthly updates) of the information (changes in the tree coverage). The section 2.3.1 of this deliverable provides general concepts for the validation, while the section **Error! Reference source not found.** proposes a continuous approach based on insights from the literature. On the one hand, the production can be a retrospective process i.e., the production takes place after the analysis period and therefore satellite images covering the period are available. In this case a single validation can be considered for the whole period. On the other hand, and in theory the production can be more frequent, for example every six-months or every quarter of a year. In this case a continuous validation can be envisaged if requested. However, this requires timely consistent availability of dataset or images. In phase 1, a single validation (QC) was carried out at the end of production and allowed to assess the correctness of the change identified in a condensed monthly version, while assessing the time lag between the event and the detection. In fact, a single-stage validation can be envisaged if it assesses the temporal accuracy of the information (i.e. is the date of detection of the change correct, or at least with an accepted delay from the ground change?). Various analyses are reported.

When operationalising this prototype, the client may be willing to have continuous validation, e.g., on a regular basis (quarter or semester), but this seems to be difficult to achieve given the unavailability of consistent datasets. A lot of time would be spent on manually generating a reference dataset. Indeed, manual validation can be time consuming and costly, especially for high frequency or high-resolution data. Setting up automated validation checks where possible, such as script-based statistical tests, quality thresholds or error flagging, could help. Automation can speed up validation, reduce human error, and help identify trends over time. One approach is still already described in section **Error! Reference source not found.**. The dataset available are listed in [AD6].

Furthermore, above-described validation uses a two-step approach with a regular sampling grid that provides crucial first information about the expected accuracies and allows calculating the number as well as allocation of the sample points within the final stratified systematic sampling approach. This is done for the different test sites over Europe (and the tropics).

However, instead of having to undertake this two-step approach for an operational roll-out over larger areas (e.g. whole of Europe) we recommend to directly use the derived accuracy values as expected accuracies for the whole of Europe to evaluate the number of sample plots and their allocation within the final stratified systematic sampling approach.

3.4 C2: Forest Disturbance mapping

3.4.1 Description

The C2 Forest Disturbance Mapping prototype aims to identifying the drivers behind forest disturbance events, specifically focusing on disturbances characterized by a decline of tree cover vitality or significant loss of tree cover density. Key disturbance agents, as defined by key stakeholders, include: (i) windthrow/ storm damage, (ii) wildfire, (iii) insect infestations (e.g., bark beetle), and (iv) human-induced disturbances (e.g., forest clearing, clear-cutting, logging, and thinning activities). Due to limited reference data, drought effects are not included in Phase 1. The detailed specifications of the C2 prototype are outlined in Table 3.

Table 3: Technical and thematic specifications for the C2 prototype

Prototype C2	
Prototype	C2 – Forest Disturbances
Thematic specifications	Forest disturbance agents' classification: (i) wind throw/ storm damage (ii) wildfire (iii) insect infestation (e.g., bark beetle) (iv) forest clearing, clear cut and logging/ thinning
Prototype sites	Phase 1: Spain (1), Central Europe (4) Phase 2: Spain (1), Central Europe (4), Sweden (5)
Temporal frequency	Monthly (Phase 1) – Annual (Phase 2)
Spatial resolution	10m
MMU	Pixel-based (options for MMU to be tested during phase 2)
Envisaged accuracy	Will be defined during Phase 2

In Phase 1, the C2 prototype provides continuous probability estimates of the likely causes of forest disturbances, using monthly forest change data from the C1 prototype as input. The prototype is highly innovative, as it is not yet part of the CLMS portfolio. Phase 1 targets monthly updates, meeting the requirement for high-frequency products for event-specific monitoring. For Phase 2, improvement of the algorithm will be targeted, and a focus on annual, aggregated classification products will be added, identifying the primary disturbance cause and assessing forest management actions (e.g. clearing post-disturbance). These products are delivered at a 10m spatial resolution across European test sites. While human-induced activities like clear-cuts and logging are detected, distinguishing forest management from natural disturbance remains challenging without additional reference data.

3.4.2 Prototype validation

3.4.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated into the validation process, consisting of a quality control of the products to look for any macroscopic errors. The product provides monthly information and the accuracy of the identified changes is assessed using Sentinel-2, Google Earth imagery, the remote sensing based National Detection System for Forest Damages (FNEWSs) for Germany containing reference data for forest disturbances (e.g. bark beetle and windthrow) and the European Forest Fire Information System (EFFIS) containing near real-time and historical information on forest fires. This quality check also allows to look at the homogeneity of the detection, i.e. are the areas of a disturbance driver homogeneous or are they mixed with other drivers.

Such a qualitative assessment must be carried out during the prototyping phase and a map of the product should be provided in the report.

3.4.2.2 Quantitative assessment

3.4.2.2.1 Sampling design

The prototype C2 deals with the identification of agents and drivers of tree cover loss or the reduction of tree vigour (drop in chlorophyll content) between different releases of the product, monthly for Phase 1 and in addition, annually for Phase 2. The validation aims to assess whether or not the identified most likely agent/ driver is correct for the known disturbance areas identified in C1.

In principle, the probability of the agent/driver is identified on wall-to-wall basis, and the thus derived map is intersected with the C1 result to show which disturbance date is related to what type of underlying disturbance driver. Therefore, the same sampling design as for C1 is used for C2.

Ancillary data such as those mentioned in the EO and in-situ data requirements report – interim version [AD3] can be used to automatically validate the agent/ driver of change. The greatest challenge for producing C2 is the lack of Europe-wide ancillary data on forest disturbance drivers including the timing when they occur, which affects both training and validation.

One part of the data is dedicated to validation purposes while another part is dedicated to classification. For example, the European Forest Fire Information System (EFFIS) provides up-to-date and reliable information on forest fires in Europe[2]. Georeferenced data can be easily downloaded on a regular basis, and this can be used to assess whether the agent "wildfire" is correctly represented in the probability map that is produced for this prototype. The remote sensing based National Detection System for Forest Damages (FNEWSs) for Germany containing reference data for forest disturbances (e.g. bark beetle and wind throw) is also used and provides an exceptionally good ancillary data situation for parts of Germany. However, it has to be ensured that training data that have been used to model the agent/driver classification is not used for validation purposes and vice-versa.

Another reference data set used to derive validation data is the European Forest Disturbance Atlas (Viana Soto and Senf, 2024) covering the period 1985-2023 with Landsat satellite data as input, but combining bark beetle infestation and wind throw into a 'bark beetle/wind complex' on an annual basis and on a lower spatial resolution (30m). It is therefore necessary to combine this information with additional visual interpretation of the sample points in order to separate these drivers.

In addition, a visual collection of sample points for logging and clear-cut events for the Central European site has been undertaken as reference data for these drivers that were lacking. For the Spanish site, reference data on forest thinnings and plantage management techniques were visually collected as well, in order to distinguish from other disturbances. The collection of this data was undertaken on the disturbance stratum of the C1 product.

[2] https://effis.jrc.ec.europa.eu/apps/effis_current_situation/

3.4.2.2.2 Response design

The interpretation of the reference data in terms of assigning the thematic class of the change agent is done by visually interpreting the data sets mentioned in the previous section. This is performed in a blind way by one interpreter. A comparison is then made between this visual analysis, which constitutes the reference, and the extraction of the product classes on the same points. Similar to C1, all sample points showing a different result compared to the reference data are double-checked by a second interpreter for plausibility. The contextual information of each pixel is crucial as the accuracy of the activity data (C1 product) also influences the quality of the C2 product, meaning that slight inaccuracies of the activity data could cause that the wrong driver is selected even though surrounding pixels are dominated by the proper driver. In this way, the approach follows the majority rule of the method described in Xu et al. (2024). Analog to the C1 approach, a consensus must be obtained with the second interpreter having higher priority. Eventually, a confusion matrix is obtained to derive the raw accuracies.

3.4.2.2.3 Analysis

As presented in section 2.3.3, User, Producer, Overall accuracies, 95%CI are calculated. Figures are available for each driver of disturbance and allow to depict the correctness of the identified driver of change.

3.4.3 Lesson learnt from Phase 1

The main challenge for this prototype, apart from the production itself, is the availability of consistent and timely data sets for the different types of drivers. It has been found that some datasets provide useful insights for a validation, but while EFFIS data are available for European countries, FNEWS data are only available for Germany. Other datasets such as the European Forest Disturbance Atlas (Viana Soto and Senf, 2024) could also be considered if they will be made available for a more recent period. However, that Atlas provides information at the most basic level and at a lower spatial resolution of 30 metres.

It was found that an annual product would certainly make more sense in relation to the time scale of some drivers and due to the greater feasibility to verify forest disturbances after a yearly analysis period. Furthermore, it must be noted that the product is not recommended to be considered at the pixel level nor as a hard classified thematic map. Instead, probability values for each agent/ driver should be shown for the time being. For phase 2 it is foreseen to test patches of pixels – instead of a single pixel analysis – to derive the main underlying disturbance agent. This would mean that the product deviates from a pixel-based approach, eventually applying an MMU.

At this stage the date of change, the homogeneity of the quality for a continuous production were not analysed during the validation process. This is something which could be considered for phase 2.

3.4.4 Operationalisation: recommendations

The method works in principle but requires a large amount of samples/reference data, which may be difficult to obtain on a European scale and will affect the validation process.

Whereas some events can be easily identified on a satellite image, others are more insidious and would need to be verified with field data or using a time series of VHR images.

In addition, regional differences in terms of the occurrence of disturbance agents must be considered, hence regional calibration is of utmost important. Which classes of disturbance agents occur in different regions, and if they can be reliably distinguished or should better be joined into fewer classes is still subject to testing.

Finally, it might be worth considering reducing the number of drivers to be analysed while concentrating only on the relevant ones in terms of forest management. In Central Europe as well as in Spain any kind of anthropogenic impact to the forest is mostly well-known whereas the most relevant drivers indicate potential outbreaks of insect infestations, which must be detected at an early stage to prevent more severe effects on the forest. Furthermore, a reduced number of drivers would allow obtaining more reliable results with higher accuracy values.

3.5 C3: Forest biomass mapping

3.5.1 Description

The C3 Biomass prototype consists of two key products: Forest Canopy Height (FCH) and Above-Ground Biomass (AGB), both targeting woody biomass (Table 4). FCH, also referred to as tree cover canopy height (in alignment with prototypes C1 and C2), enables the application of various forest definitions, providing users with flexibility in analysis and interpretation. Biomass mapping has been an active area of research for many years and is operationally conducted within the Climate Change Initiative (CCI) Biomass, though currently limited to a global spatial resolution of 100m or higher. In the CLMS context, there is significant innovation potential, as no such product is yet available in the CLMS forest portfolio. While global biomass mapping is anticipated as a future addition, Tree or Forest Canopy Height is not yet included, despite ongoing scientific developments aimed at producing large-scale tree canopy height products.

Table 4: Technical and thematic specifications for the C3 prototype

Prototype C3	
Prototype	C3 - Biomass Mapping
Thematic specifications	Above-ground woody biomass (AGB) in t/ha Forest canopy height (FCH) in meters
Prototype sites	Phase 1: Spain (1), Sweden (5) Phase 2: Spain (1), Sweden (5), additional testing: French Guiana (7)
Temporal frequency	Annual

Spatial resolution	10m
MMU	pixel-based (0.25ha and 0.5ha tested in WP2)
Envisaged accuracy	Should be on the level of the CCI Biomass results, but with a higher spatial resolution (10m) and MMU of 0.5ha oder 0.25ha

For phase 1, the derived models for both AGB and FCH, utilized a combination of optical and SAR time series, DEM-based information, and GEDI LiDAR data, land cover structures visible in Very High Resolution (VHR) reference data could be effectively depicted. However, the methods tested and implemented so far are not yet operational, indicating significant room for improvement. The models tend to saturate at higher biomass and larger tree heights, leading to underestimation of AGB and FCH. Additionally, the AGB model appears to overestimate results at lower biomass values. These initial findings highlight the need for further analysis to assess the potential for a Europe-wide application in Phase 2.

3.5.2 Prototype validation

3.5.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated into the validation process, consisting of a quality control of the products to look for any macroscopic errors and to identify areas of homogeneity and heterogeneity in the product in terms of biomass and tree height. Knowledge of the agro-ecological zones and forest types in the study area can be useful to better identify artefacts in the classification, such as tree areas that should be recorded as more or less dense and therefore with more or less biomass. Knowledge of the locations and dates of plantations, if there are any in the area, can be used to check that the product accurately reflects tree growth. This analysis simply allows the distribution of values across the landscape to be observed. The comparison of different geographical areas is also very important and makes it possible to assess the homogeneity of the product for different landscapes and possibly the need to adapt the algorithm to the agro-ecological or regional level.

An interesting comparison can also be made with existing datasets to look at the level of spatial detail, for example with the Climate Change Initiative (CCI) BIOMASS (ESA), which provides estimates of AGB for five epochs (2010, 2017, 2018, 2019 and 2020). The spatial resolution of the map products is 100 m. However, this database is also known to contain some bias, as pointed out by Araza et al. (2022).

Such a qualitative assessment must be carried out during the prototyping phase and a map of the product should be provided in the report.

3.5.2.2 Quantitative assessment

3.5.2.2.1 Sampling design

The available in-situ data will be used for validation. The sampling design is implemented to take advantage of the existing reference data set. For obvious reasons, the validation data cannot be obtained from VHR interpretation, but from existing in-situ data, such as National Forest Inventories (NFIs). The below Figure 10 displays the available NFI globally.

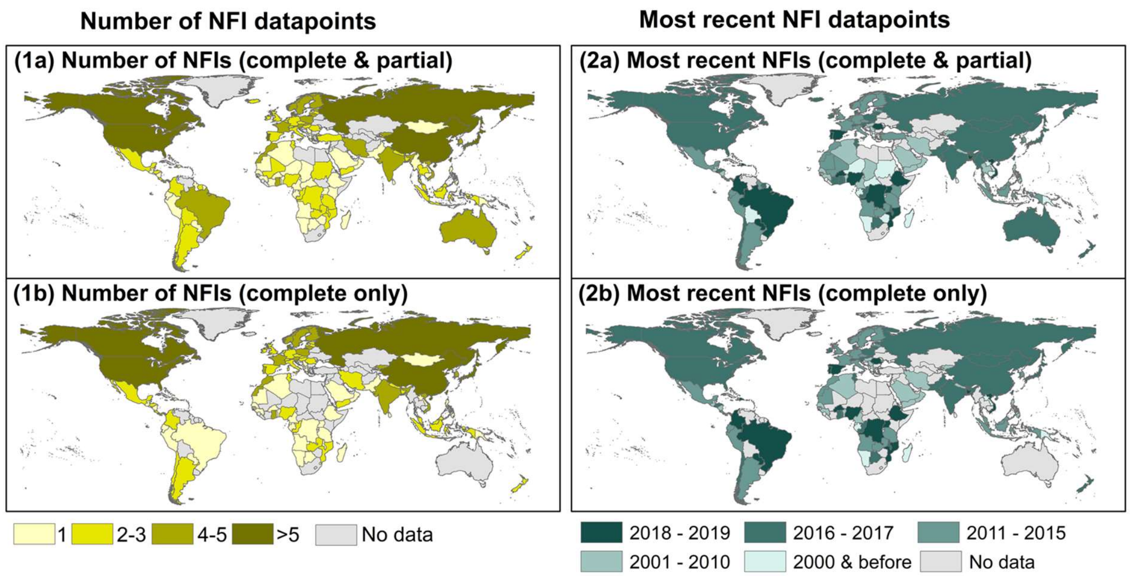


Figure 10: Country status of NFIs, recent NFIs are from the year 2010 (source: Nesha et al., 2022. Exploring characteristics of national forest inventories for integration with global space-based forest biomass data, Science of the Total Environment)

In Sweden, the reference data for both biomass and canopy height are sampled from the National Forest Inventory (NFI), which contains both above-ground biomass and mean tree height information. All sample points from the reference year 2021 can be extracted for the test site. 10% of the available points are used for validation.

In Spain, Global Ecosystem Dynamics Investigation (GEDI) Level 4A (L4A) data is downloaded and from this dataset above ground biomass (AGB) is extracted. Observations collected in power beam mode are used, incorporating all four power beams ['BEAM0000', 'BEAM0001', 'BEAM0010', 'BEAM0011']. To further refine the data, only night observations with a beam sensitivity above 0.95 are selected to minimize background noise from reflected solar radiation. For canopy height, data derived from already existing data can be used for the validation if available:

- Airborne LIDAR
- Normalized Digital Surface Model (nDSM) available locally for some test sites. nDSM represents the height of features above ground, rather than the elevation of features relative to mean sea level, or an ellipsoid. These relative height values are essential for estimating vegetation height and especially the tree height.
- The Tree Cover Height data set from Global Land Analysis and Discovery (GLAD) laboratory in the Department of Geographical Sciences at the University of Maryland. This reference dataset maps tree cover height across the globe at a 30-meter resolution and is available on Global Forest Watch (GFW) for the years 2000 and 2020. The data set defines trees as woody

vegetation of at least three meters in height and is derived from GEDI LiDAR measurements and Landsat imagery.

For the prototype sites, both a random sampling or a stratified random sampling can be applied depending on the information available. One can imagine stratifying the data to obtain an almost equal distribution of points sampled in different biomass and tree height classes to be representative of the diversity of the landscape (e.g., tree density coverage, height, slope, texture, forest age...) and inclusion weights are applied to the different strata and samples. However, it may happen that too many strata should be created to represent the different tree heigh and biomass and a too high level of complexity would be achieved – as mentioned in section 2.3.1.2. The same stratification strategy should be used for all the sites so that the results can be compared. As it can be read in Nesha et al., (2022) different type of sampling stratification is applied to produce the NFI itself in tropical countries, which has an impact on the quality of the sampling that is applied as part of EvoLand C3.

The number of samples is dependent on the number of reference points available within the area of interest.

3.5.2.2.2 Response design

Available in-situ data are used for validation. The intended data are those mentioned in the previous section and consist of the in-situ data from the National Forest Inventories, Global Ecosystem Dynamics Investigation (GEDI) Level 4A (L4A) or any in-situ or airborne LIDAR data. There is no need to perform a plausibility analysis as it is not possible to retrieve biomass or canopy height by visual interpretation.

The response design could be adjusted using the method proposed by Xu et al., (2024) i.e., validation using the sample label supplemented by the majority label within a 3 × 3 kernel neighborhood. See section 2.3.2.

The In-situ/reference data are integrated in the validation exercise only if completely spatially independent of and have not been used for the training of the classification. In order to ensure the independency between the training and the validation activities, the consortium proposes to split the available in-situ data into two sets. One which is used only for training / calibration and the other only used for the independent validation. Sampling plots for which significant changes between the field inventory date and the image acquisition date will be excluded.

3.5.2.2.3 Analytics

Biomass and Tree height are continuous products i.e. products with continuous thematic classes values. Therefore, the Root Mean Square Error (RMSE) which measures the average magnitude of the error, the Mean Absolute Error (MAE) which measures the absolute average magnitude of the errors and a linear correlation coefficient between the reference and the product will be calculated. The scatterplot obtained allows to quickly identify any overestimation and underestimation of the values from the product when compared to the reference data.

3.5.3 Lesson learnt from Phase 1

As expected, National Forest Inventories are not equally available for all test sites, and this affects both the production and validation of the prototype.

It is not always possible to select the sample points for validation specifically when they are derived from NFI data. Therefore, the validation protocol may not have the same proportion of each category of biomass and tree height.

The validation showed that the models seem to saturate at higher biomass and tree heights, thus underestimating AGB and FCH. For AGB the model also seems to overestimate the result at lower biomass values. Some differences between the results from the testing and prototyping, as well as the scattered nature of the validation plots makes the interpretation of the results not straight forward.

3.5.4 Operationalisation: recommendations

In order to operationalise such a prototype, geographical characteristics and different forest types should be taken into account when creating the dataset for both training and validation. Authors such as Araza et al. (2022), Xu et al. (2024) provide some insights into validating such a product. Araza et al. (2022) includes an uncertainty layer showing the estimated error for each pixel as a separate layer accompanying the biomass map. Spatial bias analysis and spatial error correlation models provide valuable information to both map users and producers on local to regional map errors.

3.6 C4: Cover crop type mapping

3.6.1 Description

Cover crops, also known as catch crops, are grown primarily to enhance and protect the soil rather than for harvest. They are typically planted between main crop cycles when the soil would otherwise remain bare. Common cover crops include grasses (e.g., rye, oats), legumes (e.g., clover, vetch), and brassicas (e.g., radishes, mustard). These crops offer numerous benefits, such as reducing soil erosion and nutrient leaching, increasing soil organic matter, providing natural fertilization, and enhancing biodiversity in agricultural systems. Recognized as a key practice in sustainable agriculture by the Common Agricultural Policy (CAP), effective monitoring of cover crops is essential.

Despite the establishment of various methods for mapping and monitoring main crop seasons, cover crop mapping remains less developed. Challenges include the scarcity of reliable reference data and the higher likelihood of clouds and snow during the secondary growing season, which complicates the use of optical imagery. Additionally, many cover crop seed mixes exhibit similar temporal patterns, making differentiation difficult. The objective of this prototype (Table 5) is to develop a service to map cover crop types at a 10-meter resolution annually. This will be achieved using newly collected in-situ data and a hierarchical set of interconnected Deep Learning models based on Sentinel-1 and Sentinel-2 imagery.

Table 5: Technical and thematic specifications for the C4 prototype

Prototype C4	
Prototype	Crop cover type mapping
Thematic specifications	Map cover crop types at 10m resolution utilizing novel collected in-situ data on cover crop types as well as hierarchical set of inter-connected Deep Learning models based on Sentinel-1 & 2.
Prototype sites	Phase 1: Belgium (3) Phase 2: Germany (4), Romania (6)
Temporal frequency	Yearly
Spatial resolution	10 meters
MMU	1 pixel
Envisaged accuracy	85%

In the first phase, the focus was on a test site in Belgium during the 2021 growing season. The initial map distinguishes two main groups of cover crops and a winter cereal class, based on temporal features from Sentinel-2 and VLCC layers. While grass-like cover crops were relatively well separated from other winter covers, the accuracy did not yet meet the requirements. A significant limitation for cover crop mapping is the lack of reference data. LPIS datasets from Member States rarely include cover crop information, and when they do, such as in the Flemish LPIS, the secondary crop declarations are often uncertain. This uncertainty hinders the development of a robust classification model. Additionally, defining a cover crop typology with clear separability is challenging. Currently, only two broad groups are differentiated. Cover crops often consist of seed mixes from multiple plant families, leading to confusion.

3.6.2 Validation

3.6.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated in the validation process consisting of quality control of the products to look for any macroscopic errors. The product is examined firstly in its globality. Given that this is a prototyping phase, it takes a reasonable time to go through the product in the test sites. The shape of the cover crops labelled are checked using Sentinel-2 and Google Earth imagery and also CLMS HRL VLCC and the data used for the classification.

A map of the product should be provided in the report.

3.6.2.2 Quantitative assessment

3.6.2.2.1 Sampling design

Type of sample units

Points and associated polygon/parcel are considered for the validation of the prototype C4. That means that not the precise point sample location, but rather its spatial (parcel) context is considered in the interpretation of a sample's correctness.

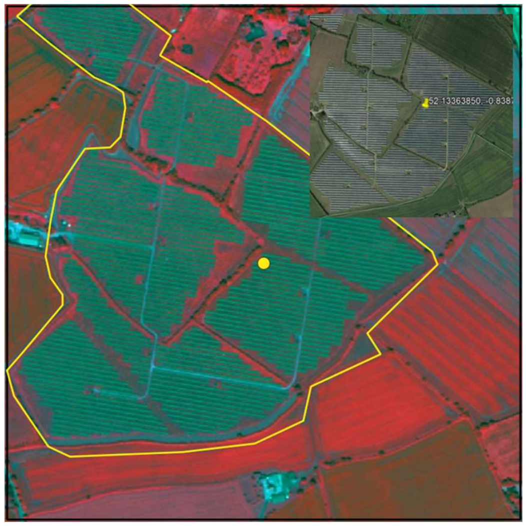


Figure 11. Type of sample units for the prototype C4

Sampling Design and Stratification

A combined approach is suggested. First, the in-situ data available will be used for the validation. Then, the in-situ data is densified with samples based on a stratified and random sampling approach. The stratification is based on the product to be validated and the cover crop classes.

The in-situ/reference data is integrated in the validation exercise only if completely spatially independent and if the data was used for the training of the classification.

3.6.2.2.2 Response design

The validation data cannot be obtained from VHR interpretation, but from existing in-situ data.

In-situ/reference data

The reference data used for validation is derived from two sources, namely:

i. LPIS Flanders (available for test site n°3)
ii. CropObserve (IIASA)

Below a brief description of the input in-situ data for the validation is provided. More details of potential reference datasets are available in the interim report related to the Existing and novel in-situ and training datasets (D2.7) [AD6].

(i) LPIS Flanders

The LPIS (Land Parcel Identification System) is a part of the Farm register system and works as a spatial representation of areas utilized by agricultural holdings which records all agricultural parcels in the Member States. It is created with two main purposes: i) clearly locate all eligible agricultural land contained within reference parcels ii) calculate their maximum eligible area (MEA). It was thus designed as the main instrument for the implementation of the Common Agricultural Policy's first pillar, whereby direct payments are made to the farmer once the land and area eligible for payments have been identified and quantified.

In most countries, LPIS contains only the main crop for a specific parcel and hence can't be used for cover crop type validation. In Flanders however, LPIS also contains farmer declaration data on secondary crops. This information is embedded in the publicly available LPIS and is used to validate the test site 3 "Belgium".

Since the parcel boundaries of the LPIS dataset do not match those in VLCC, an intersection of both parcel delineations is used to build the reference dataset. The 189 different types of crops declared as "late crops" in the LPIS are reclassified to match the mapped cover crop classes. These "late crops" declarations also include winter crops (primarily winter wheat), which are assigned to a separate class, "winter crops".

(ii) CropObserve (IIASA)

The CropObserve App is an application for collecting crop type, phenological stage, visible damage and management practices and was developed to allow anyone to observe agricultural fields anywhere. The app is focused on collecting crop type, phenological stage, visible damage and management practices. Similar locations may be revisited to track changes over time. Geo-tagged photos can also be collected to document the observation.

A local *in situ* dataset, collected with CropObserve in the framework of a Master Thesis, is used for the validation, in combination with the LPIS Flanders dataset. The *in-situ* dataset contains information on cover crop type and phenological stages for 500 parcels. The observed cover crop types are reclassified to match the mapped cover crop classes. All the parcels with *in-situ* observations are used for validation and are complemented with a subsample from the LPIS Flanders, which was discarded from the training dataset.

3.6.2.2.3 Statistical analysis

User, Producer, Overall accuracies, F-Score are calculated.

3.6.3 Lesson learnt from Phase 1

The validation process of this prototype relies on the availability and quality of ancillary datasets. High-resolution satellite imagery from Sentinel-2 or Google Earth, and land cover data like the Copernicus Land Monitoring Service High Resolution Layers (CLMS HRL) can provide information that helps verify the accuracy of the cover crop classification. However, finding precise information on cover crop types can be challenging as cover crop data often lacks the level of specificity and granularity needed to validate this prototype. For the Belgium test site, the availability of cover crop type information in the LPIS of Flanders was of first importance to calibrate and validate the prototype. The results showed that the extraction of the exact cover crop type is not yet feasible, and that grouping cover crops into broader categories (Grass-like" and "Leaf-rich & Mix," along with a separate winter cereal class) allow for meaningful validation without precise type-by-type differentiation.

During Phase 2, it will be of primary importance to test the methodology on other test sites to check its ability to map cover crop types on different geographical areas and therefore, stability of the methodology. Cover crop practices vary widely based on local soil types, climate, and farmer preferences, and may not be standardized across regions.

3.6.4 Operationalisation: recommendations

The validation process of such a product is highly dependent on the availability, quality and detail of ancillary datasets. If these datasets are limited, outdated, or lack resolution, the ability to accurately validate the cover crop mapping is restricted. Thus, the success of the validation process hinges on access to comprehensive, current, and high-resolution ancillary data, especially in prototyping phases where errors or ambiguities in classification are more likely to occur.

Publicly available information on cover crop types, like in the LPIS of Flanders, are rare, but extremely valuable for the development and operationalization of this kind of product. This could be an incentive to standardize and generalize the collection and publication of data on cover crop types and practices in Land Parcel Identification Systems.

3.7 C5: Cropland/Grassland GPP monitoring

3.7.1 Description

Information on the state and evolution of agricultural productivity is crucial for various applications, including crop yield estimation, assessing weather-related impacts on production, and measuring carbon fluxes. However, the current operational Copernicus CLMS Gross Primary Productivity (GPP) products at the European level rely on medium-resolution satellite images (e.g., Sentinel-3, MODIS), which are often insufficient for monitoring productivity dynamics at the field level. Additionally, the parameters and algorithms used are not specifically tailored to capture the productivity dynamics of agricultural lands such as cropland and grassland. Current operational products utilize the simplified Light Use Efficiency (LUE) to model GPP. Fore phase 1, the same GPP model used in the current GPP CLMS products was employed. However, the model was adjusted to estimate GPP at a 10 m resolution and to improve GPP estimations for grassland and cropland (Table 6). In the user requirements, a target accuracy of 85% was specified. However, the GPP estimation provides continuous output, making it inherently incompatible with this discrete accuracy criterion. Nonetheless, the methodology aims to deliver the most precise GPP estimation possible, utilizing the available in-situ data to its fullest potential.

Table 6: Technical and thematic specifications for the C5 prototype

Prototype C5	
Prototype	Grassland and Cropland GPP
Thematic specifications	GPP estimation at 10 m resolution over cropland and grassland locations based on Sentinel-2
Prototype sites	Phase 1: Belgium (3) Phase 2: Germany (4), Romania (6)
Temporal frequency	10 m Dekad (10-daily)
Spatial resolution	10 meters
MMU	1 pixel
Envisaged accuracy	NA

Phase 1 of the project demonstrated significant advancements in developing a high-resolution GPP product specific to cropland and grassland. The resolution was increased from 300m to 10m, allowing for observation of within-field GPP variations, meeting a key user requirement. The existing Light Use Efficiency (LUE) method was adjusted by calibrating the LUE factor specifically for crop and grassland locations, resulting in high precision, particularly for cropland. Methodological advancements were implemented in OpenEO to ensure scalability. The prototype's results were verified for consistency across 150 samples of crop and grassland spread across Europe, showing strong consistency and robustness. This consistency is crucial for the continuity of products in the CLMS portfolio.

3.7.2 Validation

3.7.2.1 Qualitative assessment

A qualitative thematic accuracy procedure has been integrated into the validation process, focusing on the quality control of the product. The calibration of the GPP product relies on a limited set of ground measurements collected over crop and grassland areas. Consequently, it is crucial to qualitatively verify the model's stability during the upscaling of the prototype. The spatial and temporal patterns of the prototype's outcomes were visually compared with those of the CGLOPS 300m GPP version for a test site located in Belgium during Phase I. This approach allowed for a qualitative assessment of the spatial and temporal consistency of the prototype, ensuring its reliability for further development and application.

3.7.2.2 Quantitative assessment

3.7.2.2.1 Sampling design

Type of sample units

Polygons were used for the validation of C5. In case of validation with reference data from Integrated Carbon Observation System (ICOS), a 30x30 m bounding box centred around each flux tower was taken (**Error! Reference source not found.**).

For the consistency check with the Copernicus Global Land Service (CGLS) 300 m GPP, a comparison was conducted using 150 random samples of 1x1 km areas spread across three test regions in Europe. Within each sample, the average GPP from both products was calculated and compared. This approach ensured a robust evaluation of spatial and temporal consistency between the two datasets across diverse regions.

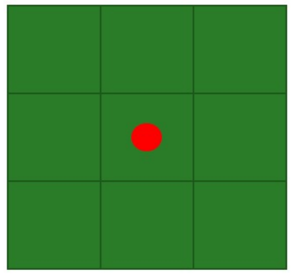


Figure 12: Sampling design over flux sites used for validation. The red dot is the flux site location, and the green squares define the Sentinel-2 pixels used for extraction.

3.7.2.2.2 Response design

As explained in section 3.7.1, GPP monitoring product is based on mathematical indices. For obvious reasons, the validation data cannot be obtained by mean of VHR interpretation but only by means of with in-situ data or consistency with similar datasets.

Evaluation against ICOS

The in-situ data available is used for the validation, namely, the ICOS sites. ICOS is a research infrastructure to quantify the greenhouse gas balance of Europe and adjacent regions. The ICOS infrastructure presents towers spread over different regions/biomes in Europe as shown in **Error! Reference source not found.**. 21 ICOS towers located in crop or grassland are selected for the prototype GPP model calibration and validation. GPP information from ICOS towers has been used for the period between 2017-2022 (note, number of available years varies from tower to tower). To ensure the independency between the training and the validation activities, the consortium proposes to proportionally split the data based on the availability of GPP information across sites, years, and months. In total 25% of the ICOS tower data is kept for the validation only.

Figure 13: ICOS infrastructure towers over different regions/biomes in Europe

The consistency of the prototype was also evaluated against the CGLS 300 m GPP dataset at a 1x1 km resolution. Although the spatial resolution of the CGLS product is lower than that of the prototype, its comprehensive coverage of Europe across multiple years makes it a valuable resource for validation. To ensure robust comparisons, a 1x1 km bounding box was selected, allowing multiple CGLS pixels to contribute to the analysis. This approach minimizes the influence of noise, which would be more pronounced if only a single pixel was used for the comparison.

3.7.2.2.3 Analysis

As presented in section 2.3.3, the RMSE, the MAE, R^2, precision is calculated. Additionally, the reference and model outputs were compared on a scatterplot, including a 1:1 line to assess whether the model tends to underestimate or overestimate GPP. These analyses provided critical insights into the model's performance and its suitability for upscaling.

3.7.3 Lesson learnt from Phase 1

The results of the Phase 1 are very promising for further improvements during Phase 2. At the moment it appears that the GPP is still overestimated during drought periods and underestimated during the off-season which should be further investigated.

3.7.4 Operationalisation: recommendations

Enhancement of the prototype still have to be done during the Phase 2, see [AD5]. Phase 1 has shown that ICOS towers and CGLS data can be used to validate the prototype, showing the importance of these kind of research infrastructures like ICOS that deliver standardised, high-precision and long-term observations which are crucial for model validation. ICOS data are available across Europe; however, the coverage for ICOS data is not uniformly distributed. This variability in ICOS data availability results in uneven regional representation, which may influence the consistency and completeness of validation efforts across different areas.

3.8 C6: Small landscape features mapping

3.8.1 Description

Small Landscape Features (SLF) are stable landscape elements that provide numerous ecological and socio-cultural functions, categorized into four main areas:

- Soil and Water Conservation: Features like grass margins, hedgerows, and ditches slow runoff, filtering pesticides and pollutants before they reach surface waters.
- Climate Protection and Adaptation: SLF play significant roles in flood and landslide mitigation, carbon storage, and sequestration.
- Biodiversity Support: SLF facilitate species movement between habitat patches, offering important habitats and ecosystem services.
- Cultural Identity: SLF structure rural landscapes, reflecting local agricultural history and providing diverse traditional landscapes across Europe, thus holding cultural value.

C6 aims to enhance the existing CLMS product, Small Woody Features (SWF), which maps hedgerows and tree patches across Europe. This will be achieved by including additional feature classes through the analysis of new Earth Observation (EO) datasets. See Table 7.

Table 7: Technical and thematic specifications for the C6 prototype

Prototype C6	
Prototype	C6 Small Landscape Features
Thematic specifications	Mapping of woody vegetated features outside forest and stable landscape features within agricultural areas (grass margin, stone walls, ditches)
Prototype sites	Phase 1: Spain (1), Belgium (3), Eastern Europe (6) Phase 2: Spain (1), Belgium (3), Eastern Europe (6)
Temporal frequency	Annual
Spatial resolution	2.5m
MMU	Not applicable
Envisaged accuracy	HRL SWF 2018 as basis: 80% UA/PA

The Small Landscape Features prototype combines supervised and unsupervised approaches, incorporates new landscape feature classes, and does not rely on existing datasets or intensive in-situ data collection, making it easily transferable to other areas. It is also computationally efficient and leverages super-resolution Sentinel-2 data, making it suitable for mapping smaller landscape elements. The existing small woody features class has been successfully extracted at an improved spatial resolution using Pleiades imagery. However, the prototype has some limitations. The new class of features is generic and does not yet differentiate specific landscape features (e.g., grass margins, ditches, stone walls) and contains commission errors that need filtering. Detecting these features from Sentinel-2 super-resolution data is a significant achievement, it implies a maximum spatial resolution of 5m. Additionally, Sentinel-2 super-resolution is not available from standard data hubs like the Copernicus Data Space Ecosystem (CDSE), necessitating substantial pre-processing before extracting landscape features.

3.8.2 Validation

During the Phase 1, the validated information was the presence/absence of detected SLF and the new class of SLF features. The studied period was 2020-2021. During Phase 2, it will be tested methods to differentiate between the new class of features which are currently generic and does not yet differentiate specific landscape features (e.g., grass margins, ditches, stone walls) and contains commission errors that need filtering. The validation will assess the presence/absence of the detection of SLF and also the correctness of the differentiation of these new elements.

3.8.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated in the validation process consisting of quality control of the products to look for any macroscopic errors. The product is examined firstly in its globality. Given that this is a prototyping phase, it takes a reasonable time to go through the product in the test sites. The correctness of the identified new SLF features is assessed with a quantitative analysis (presence/absence) as well as the stability of the method to detect all types of targeted SLF features using Sentinel-2 and Google Earth imagery; and commercial better resolved image if available. A visual check is also performed using the already existing HRL Small Woody features since the aims of this prototype is to enhance the spatial resolution of HLR Small Woody feature.

3.8.2.2 Quantitative assessment

3.8.2.2.1 Sampling design

Type of sample units

The sampling is done using sample points located at the centre of 2.5m pixel.

Sampling design and stratification

The Sampling design of the candidate C6 is a **random stratified sampling** over each test site: each AOI is divided into 3 strata, within which samples (e.g. centre of 2.5m pixel) are randomly selected. Strata were defined as:

- Area without class 1 and class 2, and also without forest mask re-used from SWF2018 - Omission strata: class 0
- SWF - commission strata: class 1
- Non-productive agricultural features (new features aggregated) - commission strata: class 2

The commission strata thus are defined by the presence of SWF and Non-productive agricultural features, the omission stratum is defined by absence of SWF and Non-productive agricultural features, both information taken from the SLF prototype raster product. As explained in Stehman & Czaplewski (1998), stratification ensures that each class is properly represented, and this sampling design allows to samples points within the small landscapes elements that are targeted in this prototype, which would be more delicate with a systematic sampling.

For each test site, 100 validation samples are distributed almost equally. During phase 1, only subset of prototype sites were produced due to VHR limited availability and samples can only be sampled within the VHR footprints. Using proportional approach would imply very few samples for rarer classes, hence this choice of having 100 in each strata. To compensate for sampling effort bias (as the 3 strata are not equal in terms of area), a weighting factor is applied to each sample, according to the strata it belongs to. During Phase 2, full extent of prototype site will be covered, so we can adjust the sampling effort to be consistent with the strata ratio and still have number of samples in each strata to would allow statistically sound results.

3.8.2.2.2 Response design

The selected samples are visually interpreted considering the applicable prototype specifications based on VHR data (i.e. pan-sharpened Pleiades at 50cm resampled at 1m spatial resolution).

The interpretation of the reference samples data is done by means of visual interpretation of VHR (Google Earth or Bing Aerial) and HR data set by **blind analysis** taking into account product specifications (MMU, MMW, class definitions, etc.). Based on the satellite imagery, the photo interpreter has the choice between several categories corresponding to the nomenclature of the mapped features (woody, ditches, grass margins, stone walls, ponds). No ancillary dataset is judged consistent enough to be used for the response design.

A comparison is then made between this visual analysis, which constitute the reference, and the extraction of the product classes on these same points, this leads to a confusion matrix to get the raw accuracies.

A second interpretation, carried out by a second interpreter as part of the **plausibility analysis**, is then done only on the sample units in disagreement with the product classification. The plausibility analysis results in revised accuracy metrics within the frame of accepted product specifications. Indeed, the blind approach described above guarantees independence from the map products but may underestimate the accuracies for complicated points when sometimes the point is located slightly out of a detected new feature perimeter, which can easily be the case with the SLF. This is resolved by the plausibility approach.

A comparison will then be made between this visual analysis, which will constitute the reference, and the extraction of map classes on these same points. A confusion matrix is then extracted, and the degrees of reliability obtained.

3.8.2.2.3 Analysis

As presented in section 2.3.3, User, Producer, Overall accuracies, 95%CI are calculated. The results presented are based on the plausibility analysis.

In addition, since the product provides information on the small elements in the landscape, it may be useful to display both the weighted matrix and the non-weighted matrix.

3.8.3 Lesson learnt from Phase 1

There are three parameters than can be assessed for such prototype: the correctness of the detection of small landscapes features (no commission), its spatial extent and the labelling of the different elements that are included in the non-productive agricultural features. Phase 1 being the first step to develop the methodology of detecting non-productive agricultural features (new features aggregated), only the correctness of the detection has been analysed, since new class of features is generic and does not yet discriminate requested landscape features (grass margins, ditches, stone walls, ponds). Spatial extent will be analysed during the phase 2 if possible, however, it will not be the main objective of the validation. The prototype has the main aim to add new feature classes and also to enhance the detection of the features.

During Phase 1, as the prototyping production started and in relation to the content of [AD5], it was planned that a stratified random systematic sampling approach will be using Land Use and Coverage Area frame Survey (LUCAS), a grid of approximatively 1,100,000 points throughout the European Union which provides diverse information on the landscape including a classification at the given points of the observed Land Use / Land Cover. PSUs and SSUs were supposed to be derived from this grid to perform the validation exercise. However, the size of the areas of interest prevented its use for the prototype validation. Indeed, the number of sample units to be selected depends on its stratum area and the number of LUCAS points intersecting with this stratum. There were not enough samples points on small linear and patchy elements of the landscapes intersecting with the LUCAS gris derived PSU/SSUs and within the available VHR 1m pixel resolution. This would have resulted in an unrepresented number of points for these classes and a complexification of the process. Instead, a random stratified sampling was done in order to directly select the samples points into 3 strata, generated with the final product. No other dataset was found to be optimal to be used as reference and to perform an automatic validation. Ideally, to get OpenStreetMap (OSM) data updated frequently would allow to get the level of details expected, however, this is not the case.

In addition, due to VHR availability and as already mentioned previously, Phase 1 only presents results over subsets of the prototype sites. Feasibility of full coverage of demonstration sites by VHR data will be assessed and performed during Phase 2.

The validation exercise allowed to identify some commissions in the final product. More details can be found in the document [AD4].

The validation shows that the target of enhancing the spatial resolution of the existing SWF2018 is also reached. The method was tested in different landscapes across Europe and proved to be stable. Already envisaged for Phase 2, the thematic classes of the change will be important to be validated so that the different aspects of the prototype can be validated.

3.8.4 Operationalisation: recommendations

Validating such product highly relies on a response design based on VHR imagery and manually interpreted labelled samples since there is no other in-situ dataset that cover a similar level of details in Europe. For an operationalisation, in the case the product is able to detect the new SLF features and to label them, two options can be envisaged.

One is to use the **same approach as in SWF 2018**, which ensures consistency with the validation approaches of previous SWF product. The sampling is done in two stages using the so-called:

1. Primary Sample Unit (PSU): 100 x 100m square segments which are considered in the interpretation of a sample unit's.
2. Secondary Sample Unit (SSU): 20-25 points which are randomly distributed within each PSU considering a 2,5x2,5m area (pixel size) for the validation analysis.

This sampling design combines systematic and stratified approaches. The design is based on the regular grid-based sampling approach, using LUCAS points to select Primary Sample Units (PSU) of 100 x 100 m squares. The stratification is based on a series of omission/commission strata based on the prototype to be validated:

- Commission stratum: SLF 1-100%
- Omission stratum: SLF 0%

The commission stratum thus is defined by the presence of SWF, the omission stratum is defined by absence of SWF, both information taken from the SLF prototype raster product. The number of sample units to be allocated to each stratum has to be calculated as a function of a class' area proportion in the geographic reference area. In case of small strata (i.e. land cover/land use (LC/LU) classes with a very low areal share), where the intersecting area between the stratum and the LUCAS points may be nominally too small to reach the required number of samples, the LUCAS grid would be replicated in order to get more sample points. Although the SLF class is a small one compared to the Non-SLF class, sub stratification was not necessary to get a sufficient number of points.

Another approach is to use a **random stratified sampling** as done for the prototype, which allows to focus more precisely on the elements detected in the final product and ensure enough points for each stratum without having to replicate the sampling several times to get enough points.

3.9 C7: Improved water bodies extent

3.9.1 Description

To better understand changes in global water surface area, the production of accurate, regular, and comprehensive water extent data is essential. Within the framework of CLMS, water area delineation is a key component. Near-real-time mapping of water bodies from satellite imagery plays a critical role in effective water management. The objective of the Improved water bodies mapping prototype (Table 8) is to enhance current CLMS products, such as the Copernicus High-Resolution Layer Water and Wetness and the Copernicus Global Land Monitoring Water Bodies. Based on the user requirements, solutions were developed starting from the existing products.

Table 8: Technical thematic specifications for the C7 prototype

Prototype C7	
Sub-prototype	Water Bodies extent
Prototype sites	Phase 1: Spain (1), Belgium (3), Sweden (5), Côte d'Ivoire (8) Phase 2: Spain (1), Belgium (3), Sweden (5), Côte d'Ivoire (8)
Description of the sub-prototype	Map water bodies extent and type of water bodies (permanent or temporary)
Expected temporal update	At least monthly
Spatial resolution	10m
MMU	0.1ha
MMW	N/A
Envisaged accuracy	>95%

The methodological developments at the end of Phase 1 resulted in several advantages. The algorithm is close to demonstration stage, meeting all main requirements with low calculation time. It maintains a native resolution of 10m on water likelihood maps and offers flexible temporality, from daily acquisitions to synthesized monthly data. It can detect new and/or ephemeral water bodies, using GSWE as a weighting factor rather than a constrained area. The current accuracy on the four test sites is higher than 97%, and the coverage is nearly global, aligned with Sentinel-2's area. Despite its strengths, some limitations need to be addressed. Currently, the algorithm relies only on Sentinel-2 data; incorporating Sentinel-1 SAR data could improve issues related to cloud cover. The exclusion mask may be too restrictive, potentially omitting some water bodies due to atmospheric strata. While static thresholds on the chosen indexes are effective, they may need tuning for global suitability. Likelihood weighting with Copernicus GLO-30 DEM slope data reduces water probability in shadow areas but may cause over-detections due to the 30m resolution compared to Sentinel-2's 10m resolution.

3.9.2 Prototype validation

During phase 1, the prototype outputs followed a validation process. To do so, the C7 prototype which is a water likelihood map, was binarized into water and non-water classes for each sites tested. The validated information is the presence or absence of water body. The studied period during the Phase 1 was 2022, and two products per year have been produced for each prototype site: winter or humid season / summer and dry season.

Belgium, Ivory Coast, Spanish and Sweden were selected as part of Phase 1 to cover a variety of landscape dynamics and complexities. The prototype area in Belgium contains rivers and water bodies which can be larger in winter compared to summer. Spanish test site has a dry climate and a lot of topography. Ivory Coast test site is a tropical zone with a very high level of clouds, and Sweden shows seasonal change with harsh winters when the landscape is partially or totally covered in ice and snow.

The following sections detail the methods used. It is also to note that this methodology might change during phase 2 to enhance the validation process.

3.9.2.1 Qualitative assessment

During the development of the prototype, qualitative assessment was part of the process: considering the low number of processed images during phase 1, time was taken to go through the product in the test sites "manually", as no reference data could be produced. Looking for macroscopic errors, this assessment allowed to improve the algorithm parameters and check its stability for the different environmental characteristics of the test sites. Such qualitative assessment was processed until the end of development for phase 1.

3.9.2.2 Quantitative assessment

3.9.2.2.1 Sampling design

To implement the validation, a **stratified random sampling design** is applied. The stratification of the area of interest is based on Global Surface Water Occurrence (GSWO) dataset which is binarized (water occurrence >10% to select all pixels that could be water) where 1 is considered as water. The sample points are extracted using 2 strata and a total of 110 samples for each test sites:

-<u>Commission (Stratum Water):</u> 30 random points out of 110 for proportional sampling in terms of water surface area in relation to the total surface area of a tile (not specific to the study zone). Based on reference dataset.

-<u>Omission (Stratum non-water)</u>: 80 points out of 110 for proportional sampling in terms of water surface area in relation to the total surface area of a tile (not specific to the study zone). Based on reference dataset.

All this process is automated, otherwise, due to the source of reference data, human supervision is necessary to correct potential misclassification errors from reference data. See next section. On this product, some pixels are classified has no data (clouds, real no data…): they are excluded from the sampling.

3.9.2.2.2 Response design

A first comparison by **blind analysis** is made from the raw sampling to estimate the proportion of true positives, true negatives and false detections. Which is followed by another check during a **plausibility analysis** to verify reference sampling points and results: in the case of water bodies mapping, confusion happens often in mixed pixels.

3.9.2.2.3 Analysis

As presented in section 2.3.3, User, Producer, Overall accuracies, F-Score, 95%CI are calculated.

3.9.3 Lesson learnt from Phase 1

As planned, a stratified random sampling of points was done. As the number of products was quite small, the validation was carried out as if each product were at T0 and were entirely validated.

Using product from EO to check the product is at risk. It means qualifying a product that is constantly evolving with static data that may not reflect these changes. Otherwise, no other dataset was found to be used as reference and to perform an automatic validation at European / world scale. Consideration was given to using VHR images, but they do not cover all the test areas.

When validating water product, misclassification is often linked to mixed pixel (representing both water and land). Two possibilities to process them: adapt the algorithm of water map estimation or keep human supervision in the validation process. Thus, the validation process helped to identify the sources of certain classification errors. Validation of the algorithm focuses on detecting the presence or absence of water without validating the extent, as this changes monthly, and the data required for accurate validation does not exist. Manual validation would be required for accurate assessment but is not applied. While it is feasible in the scope of the prototyping phases of EvoLand, it would require a dedicated budget in the context of an operationalisation. It can be noted that some papers mention the validation of global surface water extent maps using manual samples (e.g., Mukherjee et al., 2024) while other uses automated approach and/or mixed approach (e.g., CLMS Water Bodies 2014-2020, JRC[3]). In EvoLand, the validation of the ability of the algorithm to correctly map the water bodies extents is based on labelled reference data used for both training and validation of specific product status (dataset is spitted for training and validation). GSWO data, which is static but global, is used for analysis of the prototypes sites and help to validate water detections at a specific time on a specific tile, although they do not match the final products in time or space. However, validation of monthly product updates is not directly feasible. It is assumed that since the methodology remains the same, the

[3] https://land.copernicus.eu/en/technical-library/validation-report-10-daily-land-surface-temperature-v1.0/@@download/file

validity observed for one version of the algorithm will also apply to subsequent versions. In summary, the current process allows the validation of the detected extents at a particular point in time, but not their evolution over time.

The validation showed that the MMU targeted is convenient for the production and thematic classes expected while maintaining a proper level of quality.

3.9.4 Operationalization: recommendations

Two recommendations can be done from the experience acquired during phase 1.

The first one deals with a Sentinel-2 tile-by-tile approach for validation. For the evaluation of the commission and omission errors, a water/soil proportion must be given. Set as 30/80 now (i.e., 30% non-water/80% water, with 80% instead of 70% as water to densify the sampling in water areas), it could be estimated from GSWO for each tile, making this proportion more relevant from one tile to another.

The recommendation deals with the validation of different outputs in time. As presented in D3.7 [AD5], it would be interesting to follow the approach exposed in Tsendbazar et al. (2021) to validate the T1, T2, … releases from T0 release of the product to evaluate their consistency with time. Nevertheless, not enough products per year and per site were produced during Phase 1. This point will evolve during Phase 2 and the introduction of a temporary water body class, along with a denser time-series of water bodies.

3.10 C8: Continuous imperviousness monitoring

3.10.1 Description

The HRL Imperviousness product monitors soil sealing percentages and the spatial distribution of artificially sealed areas, updating every three years. While most CLMS products are updated annually, the current HRL Imperviousness layer is produced every three years. The main objective is to demonstrate the potential for continuous monitoring and incremental updates of imperviousness. The goal is to implement a prototype service at a finer 5m spatial resolution with monthly updates (Table 9). In Phase 1, the prototype was implemented every six months to consolidate and test the algorithms. Monthly updates are planned for Phase 2. The algorithms aim to maintain the quality and thematic accuracy of the products over time, ensuring temporal consistency in the time series.

Table 9: Technical thematic specifications for the C8 prototype

Prototype C8	
Sub-prototype	C8 Continuous Imperviousness Monitoring
Prototype sites	Phase 1: Southern France (2), Belgium (3), Central Europe (4) Phase 2: Southern France (2), Belgium (3), Central Europe (4), Côte d'Ivoire (8)
Description of the sub-prototype	Monitors the percentage of soil sealing and captures the spatial distribution of artificially sealed areas including the level of sealing of the soil.

Expected temporal update	Continuous (every 6 months in Phase 1; monthly in the Phase 2)
Spatial resolution	5m
MMU	1 pixel
Envisaged accuracy	HRL IMD 2018/HRL NVLCC 2021 as basis: min 90% UA/PA for urban/sealed classes

Phase 1 of the project successfully demonstrated that the developed processing chain can efficiently handle substantial volumes of input data while delivering high-quality classification results. The integration of Sentinel-2 super-resolution technology has shown significant promise, enhancing the precision and overall quality of the processed data. These advancements not only meet the initial project objectives but also lay a solid foundation for further progress in subsequent phases. Despite these successes, some issues need addressing. Errors of omission and commission are balanced across the three test sites, but commission errors appear underestimated. Differentiating rocks or bare soils remains problematic and requires attention in Phase 2. Omission errors are mainly associated with roads and isolated or scattered built-up areas. Regression and thematic accuracy results indicate an underestimation of the imperviousness degree and commission errors.

3.10.2 Validation

3.10.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated in the validation process consisting of quality control of the products to look for any macroscopic errors such as e.g., sharp built-up density alterations due to omission and commission faults; including but not limited to land cover borders, and other physical elements such as riverbeds, coastline, bare soil and rocks especially in mountain areas. The product is examined firstly in its globality using Sentinel-2 and Google Earth imagery. This qualitative assessment is also an opportunity to check whether the parameters of the algorithm could be finetuned and improved.

3.10.2.2 Quantitative assessment

During the Phase 1, the first release of the product (T0) was validated according to a full validation exercise so that the accuracy is well known and because the version T0 will be the basis for next releases of the products (T1, T2, etc.). If the results obtained during the Phase 1 for the T0 are good, the processing chain/workflow can be considered as robust and we can be confident that the same level of quality will be achieved for the next releases (T1, T2, etc.) of the prototypes as well.

3.10.2.2.1 Sampling design

The validation approach is based on a selected sample design that combines **systematic and stratified** approaches and benefits from the advantages of both of them. It is based on the LUCAS (Land Use/Cover Area frame statistical Survey) sampling approach. LUCAS corresponds to a grid of approximatively 1,100,000 points throughout the European Union where land cover or land use type is observed. Using LUCAS points ensures traceability and coherence between the different layers/iterations. LUCAS points are located every 2 km on a regular grid.

For the Continuous imperviousness monitoring prototype, a stratification is applied based on a series of omission/commission strata defined as follows:

- Commission – sealed strata: Imperviousness Degree 1-100% (IMD value = [1-100%])
- Omission -non-sealed strata: Imperviousness Degree 0% (IMD value = 0%)

It should be notice that the stratification of the areas of interest is based on the first release of the product (T0).
Each PSU corresponds to one aggregated pixel (100m). Each PSU is then associated to secondary sampling units (SSUs) corresponding to a 5x5 grid with 20m between each SSU (Figure 14).

Figure 14: Example of SSUs organised in a 5x5

For each SSU, the interpreter validates the presence of sealed or non-sealed coverage on VHR. Imperviousness degree can be extracted from such interpretation.

3.10.2.2.2 Response design

The collection of the reference data has been done on the basis of interpretation of VHR (Google Earth or Bing Aerial) and HR data set. Indeed, an experienced photo interpreter can easily classify most of the sealed based on VHR/HR images. Based on the satellite imagery, the photo interpreter has the choice between two categories, and he qualifies each SSU either as sealed or non-sealed.

The sample units are first **blindly interpreted** by the validation expert team. A second interpretation, carried out by a second interpreter as part of the **plausibility analysis**, is done only on the sample units in disagreement with product classification. The Plausibility analysis results in revised accuracy metrics. Indeed, the blind approach described above guarantees independence from the map products but may underestimate the accuracies for complicated points when sometimes several classes are possible (located at the edge of a sealed perimeter). This is resolved by the plausibility approach.

3.10.2.2.3 Analysis

Two types of analysis were conducted in phase 1:

1. Thematic accuracy
2. Imperviousness Degree (IMD) scatterplots and regression analysis

Thematic accuracy

A threshold has been applied to the density values for reference and map data to produce binary attributes of built-up for both the reference and map data layers. The threshold was set to 1) 1% and 2) 30 % with density values respectively lower than 1% and 30 % classified as 0 (non-built-up) and density values greater than or equal to 30 % classified as 1 (built-up). The 1% threshold allows the assessment of the built-up mask whereas the 30% threshold is commonly used in HRL.

Then, as presented in section 2.3.3, User, Producer, Overall accuracies, 95%CI were derived.

IMD scatterplots and regression analysis

A scatterplot is a way of displaying data against Cartesian coordinates to show and compare values for two variables within a dataset. The data is displayed as a series of points, where the x and y locations relate two variables assigned to a particular recording instance, in this case a PSU. The available measurements for each PSU are the original reference data and the mapped value from the product. In this way the relation of the reference and product information for a point can be compared to a 1:1 line which runs diagonally across the scatter plot. The closeness of a point to the point to the 1:1 line is an indication of the similarity between the reference and mapped results. The points that lie exactly on the x and y axes are related to omission and commission rather than the calibration of the IMD values themselves. See Figure 15.

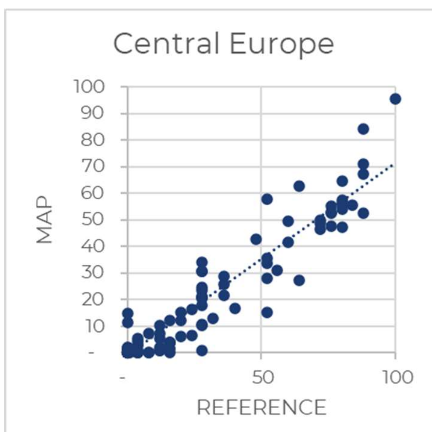


Figure 15: C8 - Example of scatterplot derived in Phase 1 for the Southern France test site

To quantitatively summarise the results displayed in the scatterplots, a linear regression analysis is performed to estimate the relationships between the reference and mapped product information. The analysis produces a coefficient of determination (R2) which is gives information about the goodness of fit of the estimated regression model. Coefficients of determination closer to 1 represent a better fit. In this case as the reference and map information are meant to represent the same information then it is useful to also consider the slope and intercept of the estimated regression model. The slope should therefore approach 1 and the intercept should be close to 0 for the required relationships. Deviations from the expected values give an indication of the correspondence of the reference and mapped imperviousness data.

3.10.3 Lesson learnt from Phase 1

During Phase 1, the validation was carried out according to the validation plan presented in D3.7, but only for the first iteration of the time series (T0). Indeed, the first release of the product (T0) was validated according to a full validation exercise so that the accuracy is well known and because the version T0 is the basis for next releases of the products (T1, T2, etc.). If the results obtained during the Phase 1 for the T0 are good, the processing workflow can be considered as robust and we can be confident that the same level of quality will be achieved for the next releases (T1, T2, etc.) of the prototypes as well.

Phase 1 allowed the quality of the prototype (T0) to be checked at two levels: 1) the degree of impermeability and thus the calibration step, and 2) the correctness of the detection of the sealed features and thus the classification process.

Phase 2 would be the opportunity to check the quality of the next release of the prototypes and also the quality of the detected changes.

The sampling design was appropriate to study the heterogeneity of the study area and the level of quality of the first release (T0). But the method presented in section 2.3.4 still needs to be tested in order to estimate the robustness of the validation approach applied to continuous products.

The results in Phase 1 showed that the main source of errors are the bare soils (quarries, crop fields without any crop development), the sands (riverbanks) the rocks and the snow, affecting mainly mountainous areas and cultivated land. This needs to be investigated and improved during the Phase 2.

During Phase 1, ground truth data was collected by photo-interpreting VHR images. This is a time-consuming task. When operationalizing this continuous prototype, a continuous validation on a monthly basis may be required. In this context, a lot of time would be spent on manually generating a reference dataset. Photo-interpretation is time-consuming and costly, so in-situ data would be required to automate validation checks for high-frequency validation. The available in-situ data sets are listed in [AD6] and could be useful for the validation of the continuous prototype.

The F-Score, which is a measure that combines user and producer accuracies into a single metric, was not derived in Phase 1 and may be added in Phase 2 to provide a balanced assessment of the classification's performance for the sealed/built-up class.

Phase 1 of the project has successfully demonstrated that the developed processing chain is capable of delivering high-quality classification results. The results indicate that the prototypes generally align with the technical specifications applied to the current Imperviousness products within the CLMS portfolio. The minimum thematic accuracy requirement of 90 % (including the CI95%). Nevertheless, the phase 1 has showed a need to enhance the quality of these prototypes. Specifically, there is a need to address and **reduce the underestimation of the imperviousness degree as well as the commission errors** that are visible through the qualitative assessment and not clearly visible in the quantitative assessment.

3.10.4 Operationalisation: recommendations

Historically in Europe, the imperviousness degree is monitored though the HRL Imperviousness layer produced every 3 years (2006, 2009, 2012, 2015 and 2018), the year 2021 being actually in production as part of the HRL Non-Vegetated Land Cover Characteristics (NVLCC). In the HLR IMD product 2018, for each of the iterations from 2009 onwards the results have contained two products: a status layer for any reference year, as well as an imperviousness density change layer between reference years. For such product, the validation was done combining systematic and stratified approaches, based on the LUCAS grid to sample PSUs and SSUs. Ancillary data such as CLC "imperviousness classes" and Open Street Map are used to derive strata, and VHR are used for interpretation of the reference information for each reference date and change values are derived from the change layers[4]. Such approach uses both ancillary data and interpretation of VHR images to derive the reference information.

In the case of prototype C8, even if the information derived is similar to the HRL and NVLCC products, the challenge lays into the temporal updates of the information. To validate the product at European scale, one approach could be to adapt the methodology of HRL IMD to another temporal frequency. In this framework, during Phase 1, when the production of the prototypes started, and in relation to the content of [AD5], it was envisaged to carry out a continuous validation of the monthly updates of the prototypes using a stratified random systematic sampling approach. However, the unavailability of a timely and consistent database made it difficult to follow such an approach. Insights from C1 are valuable for prototype C8, see section 3.3.4.

3.11 C9: Automated Land Use mapping of urban dynamics

3.11.1 Description

The WP1 user requirements highlighted a strong whish for automated annual updates of land use products like the Urban Atlas, instead of the current three-year cycle. The prototype focuses on detecting urban dynamics using a nomenclature derived from the Urban Atlas, which includes various urban and rural classes. The methodology was revised to ensure the tool can operate with or without input land-use data, aiming to detect urban dynamics using image data and potential

[4] https://land.copernicus.eu/en/technical-library/hrl-imperviousness-2018-validation-report/@@download/file

in-situ data. These detections should enable automated annual updates of change layers in the Urban Atlas, based on a Corine Land Cover-type nomenclature with classes nested at three levels. The minimum mapping units adhere to Urban Atlas rules: 0.25 ha for urban classes and 1 ha for rural classes (Table 10).

Table 10: Technical thematic specifications for the C9 prototype

Prototype C9	
Sub-prototype	C9: Automated land use mapping of urban dynamics
Prototype sites	Phase 1: Southern France (2), Belgium (3), Central Europe (4) Phase 2: Southern France (2), Belgium (3), Central Europe (4), Côte d'Ivoire (8)
Description of the sub-prototype	Urban Atlas classes level 1 to 3
Expected temporal update	Yearly
Spatial resolution	10m
MMU	0.25ha (urban areas), 1ha (rural areas)
Envisaged accuracy	85%

Phase 1 allowed to develop the first version of the prototype. Automating a process previously done by visual analysis was a significant challenge. Initially, Sentinel-2 Level 2A data was chosen as input imagery data for the prototype, applying super-resolution to achieve precise delineations of change polygons in areas lacking in-situ data. However, Sentinel-2 L3A data later proved more suitable, particularly for eliminating clouds that could be mistaken for false changes, making super-resolution unnecessary. The processing, based on regularly available Sentinel-2 data, allows for annual updates with a latency of no more than one year. Since the product is not based on any existing product, it can be produced wherever L3A data is available, potentially covering all cities, towns, and suburbs.

3.11.2 Prototype validation

During the Phase 1, the validated information was the detected changes (change / no change) in the urban areas. The studied period is 2020-2021. During Phase 2, the labelled change according to the new land use / land cover, and the labelling of the information will also be validated at level 1 and if possible, level 2 of the nomenclature. In addition, the validation will also be realised on the 2021-2022 period that will be produced.

The method below is applicable to both accuracy assessments mentioned.

3.11.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated in the validation process consisting of quality control of the products to look for any macroscopic errors. The product is examined firstly in its globality. Given that this is a prototyping phase, it takes a reasonable time to go through the product in the test sites. In addition, during the Phase 1, only the correctness of the identified change was assessed with a quantitative analysis. Therefore, the shape and the new land use / land cover labelled can also be checked during this qualitative assessment; using Sentinel-2 and Google Earth imagery; and commercial better resolved image if available. This qualitative assessment is also the occasion to check whether the parameters of the algorithm could be enhancer in a first glance.

3.11.2.2 Quantitative assessment

3.11.2.2.1 Sampling design

To implement the validation, a **stratified random systematic sampling design** is applied. A systematic approach ensures full traceability of the results compared to a random approach in which the random selection of samples is difficult to trace.

Firstly, the **stratification of the areas of interest** is based on the LC/LU classes (aggregated at Level 1) and on the change classes (aggregated in one stratum for all the changes) of the map products. A spatial buffer around the change classes is also incorporated into the stratification considering that omissions of change typically occur in close proximity to areas of mapped change, while areas mapped as stable Land Use Land Cover at larger distances from mapped change are unlikely to contain omissions. The **samples points** are extracted using to 3 strata:

- Stratum error of commission: the validation dataset is expended with samples randomly selected in the areas with detected changes. This stratum allows to look for potential over detected changes and therefore will provide insight into the user accuracy.
- Stratum high probability of error of omission: samples are randomly selected within a 100 meters buffer stratum around the identified changes in the product, assuming that some changes could be partially detected in dynamic areas or omitted in areas next to the changes well detected.
- Stratum low probability of error of omission: samples are randomly selected outside the buffer stratum, everywhere else within the area of interest (S2 Tiles). It allows to point out some potential omissions.

As mentioned in section 2.3, this approach allows to reduce the weights of omission errors when validating the product.

Important to note is that these aggregations are only valid for the sampling design, each LC/LU classes are still assessed individually (if sample units are drawn for the class considered) and thus included in the thematic reporting and confusion matrix.

The redistribution of these points by stratum is established according to the proportion of each stratum in terms of surface area in relation to the total surface area of the study zone. The number of points for each stratum will be displayed in to the description of the validation performed by the partner in charge in the next prototype delivery document.

The results will then be displayed according to a weighted confusion matrix and associated indicators, enabling the results to be analysed according to the surface representativeness of each stratum. The number of sample units is adapted to each stratum, trying to get the most reasonable weight associated, with a densification for small strata. The number of sample units per stratum should be enough to ensure a sufficient level of precision at reporting level. The minimum number of sample units per stratum should be set at ten if possible. *See section 2.3.1 for good practises on this item.*

3.11.2.2.2 Response design

The interpretation of the reference data is done by means of visual interpretation of VHR (Google Earth or Bing Aerial) and HR data set by **blind analysis** taking into account product specifications (MMU, MMW, class definitions, etc.). Indeed, an experienced photo interpreter can easily detect the change areas based on VHR/HR images comparison. Based on the satellite imagery, the photo interpreter has the choice between two categories (change / no change) and for each change he qualifies the classes at level 1 for both years. No ancillary dataset is judged consistent enough to be used for the response design.

A comparison is then made between this visual analysis, which constitute the reference, and the extraction of the product classes on these same points, this leads to a confusion matrix to get the raw accuracies.

A second interpretation, carried out by a second interpreter as part of the **plausibility analysis**, is then done only on the sample units in disagreement with the product classification. The plausibility analysis results in revised accuracy metrics within the frame of accepted product specifications. Indeed, the blind approach described above guarantees independence from the map products but may underestimate the accuracies for complicated points when sometimes the point is located slightly out of a detected change perimeter or when the notion of change is visible on VHRs (Google Earth for example) but almost indistinguishable by comparison of Sentinel images. This is resolved by the plausibility approach. The results presented in the section above are based on the plausibility analysis. In addition, since the interpretation of LC/LU is often difficult to assess in a fully blind manner, especially for points located at the edge of two classes, the polygon outline can be added to know to which side the sampling point belongs to. Therefore, even for the blind approach, the skeleton polygons, however without any thematic class information, can be used in the point interpretation.

A comparison will then be made between this visual analysis, which will constitute the reference, and the extraction of map classes on these same points. A confusion matrix is then extracted, and the degrees of reliability obtained.

During the Phase 2 of EvoLand this C9-prototype will be concentrating more on the classification of these changes to better evaluate its precision.

3.11.2.2.3 Analysis

As presented in section 2.3.3, User, Producer, Overall accuracies, F-Score, 95%CI are calculated.

In addition, since the product provides information on the dynamics of urban areas, it may be useful to display both the weighted matrix and the non-weighted matrix. Indeed, despite the introduction of a high probability omission stratum (buffer around the changes) that tends to lower the weight of omission, the area of the unchanged areas still occupied a great percentage of the area of interest.

3.11.3 Lesson learnt from Phase 1

During Phase 1, as the prototyping production started and in relation to the content of [AD5], it was planned that the stratified random systematic sampling approach will be using Land Use and Coverage Area frame Survey (LUCAS), a grid of approximatively 1,100,000 points throughout the European Union which provides diverse information on the landscape including a classification at the given points of the observed Land Use / Land Cover. These points were supposed to be sampled and used as reference dataset with the number of samples for each stratum adapted for all thematic layers, with a densification for small strata. However, several factors prevented their use for the prototype validation:

- For a given stratum, the number of sample units to be selected depends on its area and the number of LUCAS points intersecting with this stratum. For changes stratum, no points or a very few numbers of points were available, hampering the proper sampling. Indeed, only 0.12% of the detected changes intersect with LUCAS, which represents only 12 changes on 5662 (for the 3 test sites).
- If there is a willingness to use the information carried by the LUCAS database, it sometimes provides information on the Land Use and Land Cover that corresponds to an isolated part of the landscape around the sampled point on the field. Therefore, it is not always representative of the surrounding landscape on the field and does not entirely fit to the MMU that is set for the prototype (0.25 ha).
- The prototype aims to provides information on dynamics changes in urban, at a yearly frequency. However, LUCAS sampling is done every 3-4 years (2006, 2009, 2012, 2015, 2018, 2022) and the campaign can last more than 1 year long. In addition, there is no date for the information reported in LUCAS. Therefore, it seems not be a consistent database to analysis dynamics changes.

No other dataset was found to be optimal to be used as reference and to perform an automatic validation. Ideally, to get OpenStreetMap (OSM) data updated frequently would allow to get the level of details expected, however, this is not the case. It was also investigated to use of Pleaides 3D sensor as reference dataset; however, the coverage was not sufficient to allow a proper validation of the prototype.

The sampling design implemented during Phase 1 contained only 1 or 2 points per site and emphasis should be put during Phase 2 to increase it. In addition, the sampling used 3 strata including one buffer stratum of 100 meters around the commission stratum. Different sizes of this buffer should be tested, tentatively to reduce the weight of the omission stratum.

The validation exercise allowed to identify some artefacts of classification in the final product, due to the clouds and objects in the images such as parking areas and house's reflectance. More details can be found in the document [AD4].

That said, the validation revealed that the MMU targeted is convenient for the production and thematic classes expected while maintaining a proper level of quality. Already envisaged for Phase 2, the thematic classes of the change will be important to be validated so that both the correctness of the identification of change is identified, such as the correctness of the labelling of this change, if any. In this regard, label at nomenclature level 1 and if possible, level 2 will be validated as well.

3.11.4 Operationalisation: recommendations

Despite that the LUCAS database is not used for the validation of the prototype, it was still successfully employed when validating European High Resolution[5] Layer to create a grid to perform the sampling design in the past. If no dataset containing information on the land use and land cover became available when operationalization of such prototype is envisaged, LUCAS can still be used, however, the information carried by the database may not be used properly and a need for manual interpretation could be identified.

Ideally, the different updates of the European product should be updated regularly, however, if no dataset becomes available it will have to be done manually.

Deliverable [AD6] can be checked for the list of data currently available that could be used for European purposes.

3.12 C10: Continuous mapping of land surface categories

3.12.1 Description

Most operational global land cover datasets, such as the Copernicus Global Land Service (CGLS) Land Cover product at 100m resolution, NASA's MCD12Q1 dataset at 500m resolution, and the ESA Climate Change Initiative (CCI) land cover dataset at 300m resolution, classify the Earth's surface annually. While invaluable for various applications, their fixed legend systems and yearly updates do not support near-real-time monitoring of rapid or seasonal changes, such as forest loss or cropland rotations. Timely land cover information is crucial for understanding anthropogenic pressures, ecosystem states, and their evolution. Near-real-time mapping of basic, independent diagnostic categories, known as Land Surface Categories (LSC), can complement yearly land cover maps. Continuous LSC information can capture land cover dynamics and seasonal changes, supporting downstream applications like land cover and land cover change products. Other applications include mapping

[5] https://land.copernicus.eu/en/technical-library/hrl-imperviousness-2018-validation-report/@@download/file

dynamics (e.g., snow, water, vegetation cycles) and generating user-defined classification systems (e.g., ephemeral vs. persistent water bodies). See Table 11.

Table 11: Technical thematic specifications for the C10 prototype

Prototype C10	
Prototype	Continuous Mapping of Land Surface Characteristics
Thematic specifications	Land Surface Categories (LSC) are basic, independent diagnostic categories and characteristics describing the properties of the Earth's surface. This deliverable describes the LSC prototype maps, produced at 10m resolution for single scenes over the test tiles. The thematic classes mapped during the first phase include: - Woody vegetation - Herbaceous vegetation - Built-up areas - Bare - Snow/ice - Water - Thick clouds - Thin clouds - Cloud shadows
Prototype sites	Phase 1: Belgium (3), Sweden (5) Phase 2: Spain (1), Southern France (2), Belgium (3), Central Europe (4), Sweden (5), Eastern Europe (6), Cote d'Ivoire (8)
Temporal frequency	Single scenes (one scene per month for the year 2020)
Spatial resolution	10 meters
MMU	1 pixel
Envisaged accuracy	85% OA

During the first phase of the project, we developed an initial methodology for near-real-time (NRT) mapping of Land Surface Categories (LSC). Visual inspection and evaluation on the test set indicated that the LSC captured the classes with reasonable accuracy, successfully detecting changes such as alternations between bare and herbaceous vegetation and forest loss. However, there are some challenges. The probability of the woody vegetation class over sparsely vegetated areas tends to show seasonal fluctuations, and there is confusion between the built-up and bare classes. Additionally, some artifacts, such as the prediction of water over topographic shadows, are still present.

3.12.2 Validation

3.12.2.1 Qualitative assessment

Prototype sites in Belgium and Sweden were selected as part of Phase 1 to cover a variety in LSC dynamics. The prototype area in Sweden shows seasonal change in snow cover and has a significant woody cover, while the prototype area in Belgium contains more built-up and cropland areas, with an alternation between bare and herbaceous vegetation.

A qualitative thematic accuracy procedure is integrated in the validation process consisting of quality control of the products to look for any macroscopic errors. The product is examined firstly in its globality. It delivers monthly data and the accuracy of the identified LSC is evaluated using Sentinel-2 and Google Earth imagery. The consistency and the homogeneity of the product is checked for several months, e.g., are the temporal changes in the LSC classes correct or are they caused by confusion with other classes? Such visual inspections are particularly crucial in dynamic areas, such as agricultural lands, where seasonal transitions—bare soil, growth stages, and harvest—occur within the year; and the geographical specificities of the test sites are checked.

3.12.2.2 Quantitative assessment

3.12.2.2.1 Sampling design

Type of sample units

Patches (polygons) are considered for the validation of the prototype C10.

Sampling Design

Manually collecting annotations from single scenes and different dates is very costly and time consuming, therefore an automated workflow combining auxiliary data to collect the annotations is preferred. Specifically, the WorldCover 2021 labels to derive the most likely LSC label is used. The In-situ/reference data is integrated in the validation exercise only if completely spatially independent of and have not been used for the training of the classification.

10% of the annotated patches are used for the validation, and a **random sampling** is applied so that each category of LSC is assessed.

3.12.2.2.2 Response design

As previously stated, the response design employs automatically labelled patch samples. The ideal scenario would be to use a timely and consistent ancillary dataset to generate a reliable reference dataset. However, such a dataset is not yet available. As an alternative, manual labelling of patches could be considered. However, given the product's objective of mapping Land Surface Cover (LSC) on a near-real-time (NRT) basis, this approach would be unfeasible due to the high costs and time demands involved.

3.12.2.2.3 Analysis

As presented in section 2.3.3, User, Producer, Overall accuracies and F-Score are calculated. Monthly or seasonal analyses are feasible; however, the sample size within each stratum may be insufficient to ensure statistical robustness. Displaying both weighted and non-weighted matrices can provide a more comprehensive view of product quality.

3.12.3 Lesson learnt from Phase 1

During Phase 1, when the production of the prototypes started, and in relation to the content of [AD5], it was envisaged to carry out a continuous validation of the monthly updates of the prototypes using a stratified random systematic sampling approach as described by Tsendbazar et al. (2021) and presented in section **Error! Reference source not found.**. However, the unavailability of a timely and consistent database made it difficult to follow such an approach. Instead, the response design employed automatically labelled patch samples. The validation showed that during some months, the spectral signature of herbaceous and woody vegetation is similar. This causes confusion between the vegetation classes and hence also seasonal fluctuations. Yet, seasonal change in woody vegetation cover is still noticeable. In addition, there is sometimes still some confusion between bare and built-up areas. It will be considered to merge those two classes as change in built-up areas is typically a relatively slow process, for which annual change layers likely provide sufficient information. Finally, some commonly occurring artefacts (prediction of water over topographic shadows or prediction of snow of undetected clouds) need to be handled. In conclusion, LSC capture the classes with reasonable accuracy. Changes, such as alternations between bare and herbaceous vegetation or forest loss were detected.

3.12.4 Operationalisation: recommendations

The production on two test sites during Phase 1 have shown that it is possible to automatically label patches of information to create a reference dataset useful both for calibration and validation. Such labelling allows to get a robust network for ground truth data collection across various regions, land cover types, and seasons which is essential for validation. Otherwise, the collection of reference dataset could also include partnerships with national agencies, local land managers, and citizen science initiatives, and time allocated to this task. For unbiased validation, stratified random sampling design should be applied across different LSC types and geographies, along with weighted accuracy matrices.

3.13 C11: On-demand land cover mapping

The purpose of the on-demand land cover mapping prototype is to demonstrate an on-demand land cover service. This will be done by showcasing a tailored land cover product with customized classes (e.g. crop types, leaf types, plantations) for different spatial resolutions (100m, 10m, 5m) and observation periods. Hence, the focus of the prototype is to develop a novel IT system and pipeline demonstrator, rather than developing an actual product.

3.13.1 Description

A wide variety of ground-truth datasets for land cover classification exist in private and public data storage spaces and are used to train land cover classification models. Oftentimes, these datasets are subjected to privacy regulations or users would like to use these datasets within their own applications without sharing the data. However, building and training a land cover classification model is an intensive process requiring specific skills and experience which not all potential users might have. Furthermore, a model trained and conducting inference on satellite data needs to be able to extract, download, and process large quantities of data, which could also be difficult for users who may not have access to the network resources required to conduct these tasks. The C11 prototype aims to address the issues above by demonstrating a cloud-based model training and inference pipeline that allows users to upload reference training data, extract training data from the CDSE backend, train a Machine Learning classification model, and choose the area, time period, and spatial resolution of the final classification results. This creates a classification map tailored to the user's needs and requirements, while allowing them to maintain the privacy of their reference data. This prototype (Table 12) will seeks to demonstrate the application of the General Framework for Mapping (GFMAP) package that standardizes and simplifies the usage of many common processes and functions used in openEO by building the model training and inference pipeline using GFMAP.

Table 12: Technical thematic specifications for the C11 prototype

Prototype C11	
Prototype	On-demand Land Cover Mapping
Thematic specifications	On-demand tree species classification, currently implemented on training data and predictions in Flanders. The tree genus classes mapped in the current prototype are as follows: - No tree - Pinus - Fagus - Quercus - Picea - Other Broadleaved - Other Needleleaved
Prototype sites	Phase 1: Belgium (3) Phase 2: Multiple sites across Europe, TBD
Temporal frequency	Single scene
Spatial resolution	10 meters
MMU	1 pixel
Envisaged accuracy	Modelling accuracy will depend on user input training data

During the first phase of prototype development, we built and tested various aspects of the pipeline, including training data preparation, feature extraction, model training and upload, and inference in openEO. We successfully demonstrated the feasibility of the prototype and delivered classified tree species maps covering areas up to entire Sentinel-2 tiles. However, we identified some inefficiencies related to processing and costs that need to be addressed in the next phase. Despite these challenges, the pipeline is functional, and further refinements will enhance its usability and customization for end users.

3.13.2 Validation

3.13.2.1 Qualitative assessment

A qualitative thematic accuracy procedure is integrated in the validation process consisting of quality control of the products to look for any macroscopic errors. The product is examined firstly in its globality.

The accuracy of the identified tree genus level is evaluated using Sentinel-2 and Google Earth imagery. To comprehensively evaluate the stability and robustness of the methodology used in this prototype, it is ideal to sample data from diverse geographical areas.

3.13.2.2 Quantitative assessment

3.13.2.2.1 Sampling design

Type of sample units

Samples points are used.

Stratification and sampling design

The in-situ/reference data is integrated in the validation exercise only if completely spatially independent of and have not been used for the training of the classification. Then, the in-situ data is densified with samples based on a **stratified random sampling approach**. The stratification is based on the tree genus level from the reference dataset and using strata corresponding to the different tree genus levels presents in the area. 25% of the total number of training points used to train the model and distributed across six classes and throughout the test sites can be used. After splitting into training and test datasets, the model is trained and validated to derive accuracy metrics.

The redistribution of these points by stratum is established according to the proportion of each stratum in terms of surface area in relation to the total surface area of the study zone. The number of sample units is adapted to each stratum, trying to get the most reasonable weight associated, with a densification for small strata. The number of sample units per stratum should be enough to ensure a sufficient level of precision at reporting level. The minimum number of sample units per stratum should be set at 10 if possible. *See section 2.3.1 for good practises on this item.*

The results are then displayed according to a weighted confusion matrix and associated indicators, enabling the results to be analysed according to the surface representativeness of each stratum.

3.13.2.2.2 Response design

The collection of the reference data cannot be done based on the interpretation of VHR and HR data sets given the level of information expected. Therefore, in-situ data is used.

In-situ/reference data

Depending on the location of the prototype site, training and validation features are extracted from a collection of ground-truth tree species local datasets. The National Forest Inventory also provides in-situ observations of dominant tree species. This in-situ data can be used to train regional or global dominant tree species models. As part of the prototyping Phase 1, tree species training data was gathered from internal training datasets in Flanders, Belgium. Test datasets used for training were private, however in future phases open access datasets will be used for validation.

Once gathered, the reference dataset is split between validation and production to ensure independence. Although some training data was provided at the tree species level, after initial testing it is determined that tree species were too similar and resulted in too much confusion in the model. Thus, all model training and inference was conducted on the training datasets aggregated at the tree genus level.

3.13.2.2.3 Analysis

As presented in section 2.3.3, User, Producer, Overall accuracies and F-Score are calculated.

3.13.3 Lesson learnt from Phase 1

The validation of this product is dependent on the availability of a suitable reference dataset, which must be sufficiently detailed and dense. Stratified systematic sampling is not a feasible option in this case, as the data are not distributed homogeneously across the territory. Therefore, a stratified random sampling approach is recommended, with the aim of sampling points from all mapped classes.

However, an imbalance in available training data for certain classes—such as *Quercus* and other broadleaved species—has been observed. Such imbalances may also arise when users upload their own reference training data, potentially leading to classification inaccuracies. Additionally, while some training data included information at the tree species level, initial testing showed that closely related species caused excessive confusion within the model, impacting classification reliability.

The accuracy of the product is closely tied to the quality and granularity of the in-situ data used for both training and validation. Users should be aware that variations in in-situ data availability may affect the product's performance, and this consideration should be taken into account when interpreting results. Accuracy is reported in the prototype via basic model accuracy measures and a confusion matrix. Accuracy will also depend on the availability of training data within the zone of interest selected by the user. All input datasets are available across Europe, however if additional input datasets are added during Phase 2, their availability will need to be verified. The prototype is ready to test with other types of land cover training datasets, which will contribute to the robustness of the prototype and the further development of potential applications. The next phase should focus on further operationalising the prototype in order to make it more accessible to the end users.

3.13.4 Operationalisation: recommendations

It was possible to demonstrate the feasibility of this prototype and deliver classified tree species maps covering areas up to entire S2 tiles. Tree species datasets are not always available for each European country, and as the prototype is developed to work with a variety of input datasets, the next steps for operationalisation of the prototype will focus on the testing of other input land cover datasets. Furthermore, next steps for operationalisation should focus on improving user access to the prototype, likely through integration into an existing web interface with upload/download options for user-provided training data and results, respectively. Additional checks on quality of user-input datasets can also be developed as the prototype is tested with a variety of different input datasets.

4 Conclusion

This document, which is the final version of the Prototype Validation Plan for the EvoLand project, concludes the prototype validation process.

The interim version of the validation plan (D3.7), delivered in December 2023, established the foundation for the comprehensive testing conducted during Phase 1. The current deliverable (D3.8) provides an update of this plan considering last available descriptions of the prototypes and expected work in to carry out during Phase 2. The validation protocols for the prototypes are outlined and are refined according to the outcomes of the Phase 1. This ensures a robust validation process but also allows us to leverage valuable insights gained from the initial prototype validation.

It is important to note that Phase 2 has only just begun, meaning the final product descriptions are likely to evolve and will be finalized by the end of this phase. For several prototypes, this includes potential changes to class descriptions, input and validation datasets, target quality metrics, and the detailed validation process.

For all the prototypes, the validation plan is subdivided into three main steps:

i. Stratification and sampling design
ii. Response design (visual interpretation of reference samples on VHR data (QA/QC breakpoint) or in-situ data/ancillary data collection; and
iii. Statistical analysis, accuracy assessments (Producer's Accuracy, User's Accuracy, Overall Accuracy, Kappa, F-Score) and reporting at the various required levels.

We are committed to adopting the most pragmatic validation approach for the prototypes. This validation plan, which includes test methods and lessons learned, aims to serve as the foundation for a comprehensive and operational service. Additionally, it seeks to provide recommendations for the validation process of fully delivered products across Europe.

The 11 prototypes are grouped into 5 main themes (forest, agriculture, water, urban, general land cover), have different categories (status, change), with different representation of the information (continuous products, Boolean products) and different temporal frequencies (near real time, monthly, yearly). Each combination of these factors introduces unique challenges and considerations, shaping both the validation process and the reliability of results. Therefore, validation has been adapted to all prototypes according to the best practises, the data available and the time allowed to this task.

Some prototypes enhance the spatial resolution of an existing CLMS service already produced at the European scale. In such cases, validation concepts can be leveraged to support the validation of the prototype's operationalization. Other provides prototypes bring new thematic specificities and/or higher temporal frequency of the information, in this case, and as seen in the current document, more efforts must be put forward for the validation to be adapted to the specifies.

Validating at European scale remains challenging, one of the main ones being the gathering of timely and consistent reference dataset used for the response design of the validation process. Availability of such dataset impacts the possible sampling design, change the inclusion probability of the samples and the results of the assessment. It remains a trade-off between the budget that can be engaged to support manual interpretation of samples and the availability of timely and consistent in-situ/ancillary dataset to create a reference.

In this context, the process of validating a change product has special considerations which make it different from that of an individual status characterization, and there is also a difference between validating a product produced with one methodology or derived independently with several methods because for example, different production sites have been involved. If the initial products t0 are of inferior quality, then the validation exercise could end up being an investigation of errors found in the input land cover characterizations, not a measure of actual change. In the two cases, the knowledge of the quality of the initial t0 product is essential to ensure that the methodology can generate the excepted specification.

- When the updates of t0 are produced dependently, they should maintain a similar quality, assuming the method is stable. For products with a temporal frequency of less than a year (e.g., 10-day or monthly intervals), the methodology should account for different types of phenology. In such cases, temporal frequency becomes a key parameter when selecting the methodology, as the validation effort must be employed efficiently. Here the approach proposed by Tsendbazar can be a solution to provide updated figures with a limited effort when compared to a classical validation of all the realises.
- In the other hands, in case the different updates of a product are produced independently using different methodologies, it is recommended to validated different releases as if they were t0 to ensure the homogeneity and stability of both methods and therefore accuracy to map the same product.

For such change prototypes, during Phase 1, ground truth data was collected using ancillary datasets when available, automatically labelled samples or by photo-interpreting VHR images.

In-situ data would be required to automate validation checks for high-frequency validation, the available in-situ data sets are listed in [AD6], and could be useful for the validation of the continuous prototype. However, using ground or reference data may contain errors and thus misclassification does not always indicate a mistake in the classification used to derive the map.

This being said, for most of the prototypes it was possible to follow a strict statistical sound approach, and to adapt the validation concepts to the prototypes. However, some adaptations were required for some prototypes with a sampling design strictly restricted by the availability of areas covered by VHR data (e.g. C6), some prototypes manually selected samples to determine the product's performance (e.g. C1) and for some prototypes it was not possible to validate all the releases of produced information (continuous products e.g., C1, C7, C8): validation of monthly product updates is not directly feasible and it was assumed that since the methodology remains the same between the releases, the validity observed for one version of the algorithm will also apply to subsequent versions.

Prototypes lacking associated accuracy data remain untested hypotheses. Assessing their quality is a crucial aspect of the prototyping phases, enabling informed decisions on which prototypes can be further operationalized at a higher level. In this context, prototype validation successfully determined whether the product specifications were both useful and feasible given the available data. For all prototypes it provided useful insights to enhance the methodology during Phase 2 and also to adapt the temporal frequency (e.g. C2), the scale of analysis (e.g. C2), or the thematic classes (e.g. C4) for example.

Finally, a standardized prototype validation framework has been developed so that all the prototypes get the same level of details when reporting the validation that will be undertaken during Phase 2.

5 References

Allen, J.D., 1990. A Look at the Remote Sensing Applications Program of the National Agricultural. J. Off. Stat. 6, 393–409.

Araza et al. (2022), A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps, Remote Sensing of Environment, 272, 112917, available at https://doi-org.ezproxy.library.wur.nl/10.1016/j.rse.2022.112917

Arevalo, P., Woodcock, C.E., Olofsson, P., 2019. Continuous monitoring of land change activities and post-disturbance dynamics from Landsat time series: a test methodology for REDD+ reporting. Remote Sens. Environ. https://doi.org/10.1016/j.rse. 2019.01.013. (in press).

Banko, G. "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory," International Institute for Applied Systems Analysis, Laxenburg, 1998.

Bellhouse, D.R., 1988. 6 Systematic sampling, in: Sampling, Handbook of Statistics. Elsevier, pp. 125–145.

Brooks, E.B., Wynne, R.H., Thomas, V.A., Blinn, C.E., & Coulston, J.W. (2014). On-the-Fly Massively Multitemporal Change Detection Using Statistical Quality Control Charts and Landsat Data. IEEE Transactions on Geoscience and Remote Sensing, 52(6), 3316–3332. https://doi.org/10.1109/TGRS.2013.2272545

Büttner, G., Kosztra, B., Maucha, G., Pataki, R., 2012. Implementation and achievements of CLC2006. ETC-LUSI, EEA.

Cochran, W.G., 1977. Sampling Techniques (3th Edition).

Congalton, R. G., & Green, K. (2009). Assessing the Accuracy of Remotely Sensed Data: Principles and Practices (2nd ed.). Taylor & Francis Group.Fichet 2014

Duncanson, L. et al. (2021). Aboveground Woody Biomass Product Validation Good Practices Protocol.

FAO, 2006. Global Forest Resources Assessment 2005 – progress towards sustainable forest management. FAO Forestry Paper No. 147. Rome.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850–53.

Lowell, K., Jaton, A., 2000. Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources. CRC Press.

Mukherjee, R., Policelli, F., Wang, R., Tellman, B., Sharma, P., Zhang, Z., & Giezendanner, J. (2023). A globally sampled high-resolution hand-labeled validation dataset for evaluating surface water extent maps. Earth System Science Data Discuss [Preprint],. Retrieved from https://doi.org/10.5194/essd-2023-168

Mullissa, A.; Vollrath, A.; Odongo-Braun, C.; Slagter, B.; Balling, J.; Gou, Y.; Gorelick, N.; Reiche, J. Sentinel-1 SAR Backscatter Analysis-Ready-Data Preparation in Google Earth Engine. Remote Sens. 2021, 13, 1954. https://doi.org/10.3390/rs13101954

Nesha, K., Herold, M., De Sy, V., De Bruin, S., Araza, A., Málaga, N., Gamarra, J. G. P., Hergoualc'h, K., Pekkarinen, A., Ramirez, C., Morales-hidalgo, D., & Tavani, R. (2022). Exploring characteristics of national forest inventories for integration with global space-based forest biomass data. *Science of the Total Environment*, *850*, Article 157788. https://doi.org/10.1016/j.scitotenv.2022.157788

Olofsson et al. (2020), Mitigating the effects of omission errors on area and area change estimates, Remote Sensing of Environment, 236, 111492, available at 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 Sens. Environ. 148, 42–57.

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 Sens. Environ. 129, 122-131, https://doi.org/10.1016/j.rse.2012.10.031

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., ... & Hofton, M. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment*, *253*, 112165.

Radoux, J., & Bogaert, P. (2017). Good practices for object-based accuracy assessment. Remote Sensing, 9(7), 646.

Radoux, J., & Bogaert, P. (2020). About the pitfall of erroneous validation data in the estimation of confusion matrices. Remote Sensing, 12(24), 4128.

Reinosch, E., Backa, J., Adler, P., Deutscher, J., Eisnecker, P., Hoffmann, K., Langner, N., Puhm, M., Rüetschi, M., Straub, C. and Waser, L.T. (2024). Detailed validation of large-scale Sentinel-2-based forest disturbance maps across Germany. Forestry: An International Journal of Forest Research, p.cpae038.

Sannier, C.A.D., Taylor, J.C., Plessis, W.D., Campbell, K., 1998. Real-time vegetation monitoring with NOAA-AVHRR in Southern Africa for wildlife management and food security assessment. Int. J. Remote Sens. 19, 621–639. https://doi.org/10.1080/014311698215892

Santoro, M. et al. (2021). Remote Sensing of Environment, 253, 112235.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling. Springer Science & Business Media.

Selkowitz, D.J., Stehman, S.V. (2011), Thematic accuracy of the National Land Cover Database (NLCD) 2001 land cover for Alaska, Remote Sens. Environ. 115, 1401–1407, https://doi.org/10.1016/j.rse.2011.01.020

Stehman (2014) Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes, Int. J. Remote Sens. 35 4923–39, available at https://www.tandfonline.com/doi/full/10.1080/01431161.2014.930207

Stehman, S. V, & Foody, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment, 231, 111199.

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. International Journal of Remote Sensing, 30, 5243–5272.

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(12), 3044-3055.

Stehman, S.V., 2012. Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. Remote Sens. Lett. 3, 111–120. https://doi.org/10.1080/01431161.2010.541950

Stehman, S.V., Foody, G.M., others, 2009. Accuracy assessment, in: The SAGE Handbook of Remote Sensing. Sage London, pp. 297–309.

Szantoi, Z., Jaffrain, G., Gallaun, H., Bielski, C., Ruf, K., Lupi, A., ... & Brink, A. (2021). Quality assurance and assessment framework for land cover maps validation in the Copernicus Hot Spot Monitoring activity. European Journal of Remote Sensing, 54(1), 538-557.

Taylor, C., 1997. Regional crop inventories in Europe assisted by remote sensing : 1988-1993 : synthesis report of the MARS Project - action 1 - Kings Norton Library (No. EUR 17319 EN). JRC, European Commission.

Tsendbazar, N., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., 2021. Towards operational validation of annual global land cover maps. Remote Sens. Environ. 266, 112686.

Wack, Hirschmugl, Koboltschnig, Schmitt, 2012. Validation Methods for REDD, Technical Note for the ESA GSE Forest Monitoring REDD Extension services project.

Westfall, J.A., Patterson, P.L., Coulston, J.W., 2011. Post-stratified estimation: within-strata and total sample size recommendations. Can. J. For. Res. 41, 1130–1139. https://doi.org/10.1139/x11-031

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. Volume 311. 2024, 114316, ISSN 0034-4257. https://doi.org/10.1016/j.rse.2024.114316 https://www.sciencedirect.com/science/article/pii/S0034425724003341

Zhang, H.K., Roy, D.P., Yan, L., Li, Z., Huang, H., Vermote, E., Skakun, S., Roger, J.-C., 2018. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. RSEnv 215, 482–494. https://doi.org/10.1016/j.rse.2018.04.031

Evoland
LAND MONITORING EVOLUTION

EvoLand Consortium