

Dear Reviewer, we appreciate all the efforts and valuable suggestions for manuscript improvements. Below you can find responds to all your comments:

Comment 1:

I recommend that the authors revise their terminology and adapt terms recommended by good practice guidance on land cover and change accuracy assessment by Tyukavina et al 2025. In particular, Section 2.1 of this guideline provides valuable guidance on definition and terminology. I recommend this because of some inconsistencies in this manuscript: samples – in geospatial applications, a sample means a collection of sampling units or sampling points. The singular word “sample” along with “sample units” should be used in this manuscript, the latter for when referring to a number of sample units (so not “samples”). Accordingly, I suggest the title “LULC validation dataset of Europe 2017.”

Response 1:

The title was modified according to the suggestion, and now it is “Validation dataset for the Land Cover Map of Europe 2017”. Throughout the manuscript, we now distinguish the reference dataset from individual sample units and use “primary sampling unit (PSU)” for Sentinel-2 tiles and “secondary sampling unit” for the 10 m Sentinel-2 pixel. We have removed the alternating use of “validation sites”, “tiles”, “granules”, and “samples” where it created ambiguity. The terminology has also been aligned more closely with the CEOS good-practice framework cited by Tyukavina et al. (2025).

Comment 2:

A quality (thematic accuracy) of the labels was not mentioned in this manuscript. Please provide information on this if it is available. #48 mentions further verification by a team of independent specialists. Please provide more information, as quality information is critical for reference datasets.

Response 2:

We agree that thematic reliability must be distinguished from technical data quality. Section 2.3 now describes the response-design screening, visual interpretation, use of ancillary evidence, rejection of ambiguous/heterogeneous locations, and independent verification. Section 4 now explicitly states that the 7% rejection rate is a response-design screening rate and is not a thematic accuracy estimate. Each sample unit was verified twice, but independent validators. They have agreed in 93% of cases. Each disagreement caused the removal of sample unit from the final dataset.

Comment 3:

Although a proportional allocation was used as the basis for the two stages, small countries and minority land cover types were treated separately. This meant all sample units do not have the same inclusion probability as in simple random sampling. In the usage section, it is recommended to provide information about the estimators (accuracy assessment and area estimation formulas). In addition, as records of the reference data, the land use and land cover map/stratification labels are advised to be included, since they were used for stratification, and whether this stratification map is openly accessible to users (for interested users to calculate the inclusion probability). In general, for a stratified sample, it is necessary to have the strata weight/area, to allow area-weighted accuracy estimates by considering unequal inclusion probabilities.

Response 3:

We agree. We now explicitly state that the design is not self-weighting because the first-stage tile selection and second-stage stratified allocation do not give all sample units the same inclusion probability. We also identify the three sampling layers: the Sentinel-2 MGRS tile grid, country boundaries, and Corine Land Cover (CLC). All of them are openly accessible and can be reconstructed following the procedure described in the manuscript.

Comment 4:

Heterogenous sample points were removed from the dataset. Please make this clear in the abstract, as there is a limitation in representing heterogeneous landscapes, which can be significant in Europe.

Response 4:

We agree and have made this limitation explicit in the Abstract, Section 2.3, Section 4, and Usage Notes. Candidate locations were retained only when the 3 × 3 pixel neighbourhood supported a single definitive class.

Comment 5:

11 – two-step stratified random sampling – later it is mentioned two stage cluster sampling. Which one is correct? Please adopt a consistent term

Response 5:

We agree. The terminology is now consistent: “two-stage sampling design with clustered first-stage selection and stratified second-stage sampling”. Sentinel-2 tiles are PSUs selected at the first stage; 10 m pixels are secondary sampling units selected within CLC strata at the second stage.

Comment 6:

#74 – What layers were selected and why?

Response 6:

We have added the sampling layers explicitly. The Sentinel-2 MGRS tile grid defines the PSUs; country boundaries define the first-stage country strata and target population; and CORINE Land Cover provides the second-stage land-cover strata. The text explains the role of each layer and notes the open availability of CLC through the Copernicus Land Monitoring Service.

Comment 7:

#67 – What does the generalization criteria mean here?

Response 7:

These part describes general approach which is following validation planning. It is crucial e.g. in validation of CLC products where various minimum mapping units are applied and generalizations is present. In case of dataset presented in this manuscript, no generalization steps was performed.

Comment 8:

Figure 4: Please indicate what the “other” class is. This is not present in Table 2.

Response 8:

Meaning of “other class” is described in #105. This class was combined from all CLC classes which could not directly correspond to S2GLC classes.

Comment 9:

Sometimes, validation sites, tiles were mentioned. I understand this is PSU. Please make this consistent.

Response 9:

We now consistently use “primary sampling unit (PSU; Sentinel-2 tile)” and “secondary sampling unit (10 m pixel)”. The Table 1 and Table 3 captions and the Figure 2 caption have also been revised. “Base PSU” and “added PSU” are defined explicitly.

Comment 10:

#145-150 – Please clarify if one aggregated image per year was used for interpretation or individual chips of different dates.

Response 10:

We clarified that the reference imagery was intended to correspond to the 2017 reference period and to be independent of the S2GLC production data. If single image was not sufficient for LCLU class identification, images from additional dates were used.

Dear Reviewer, we appreciate all the efforts and valuable suggestions for manuscript improvements. Below you can find responds to all your comments:

Comment 1:

Expert interpretation and inter-interpreter consistency are insufficiently documented. The study states that candidate samples were visually interpreted by experienced and trained interpreters and were subsequently verified by independent specialists. It also states that uncertain samples were removed. However, the manuscript does not provide sufficient information about the actual interpretation and verification process. The study should clarify the following points:

How many interpreters participated?

Response:

There were 2 independent interpreters

Did each interpreter classify a separate subset of samples, or were some samples interpreted independently by multiple experts?

Response:

Each of them has classified all samples

Was the independent verification applied to all retained samples or only to a subset?

Response:

It was applied to all samples.

What criteria were used to accept, revise, or reject an initial label? How were disagreements between interpreters resolved?

Response:

Disagreement between interpreters resulted in automatic rejection of sample

Was a senior expert responsible for final adjudication?

Response:

No

Most importantly, the manuscript does not report any quantitative measure of agreement among experts. Since the reliability of the dataset depends strongly on visual interpretation, the authors should provide an inter-interpreter consistency assessment. This could include the percentage agreement and an appropriate agreement statistic, such as Cohen's kappa or Fleiss' kappa, based on a common subset classified independently by multiple interpreters.

Response:

Unfortunately, inter-interpreter statistic cannot be reconstructed at this stage.

Comment 2:

The available attributes of data is limited. In the table 4, the data only has three attributes: S2GLC, TILE, and NAME_ENG. These fields allow basic use, but they provide limited traceability and constrain reproducibility. For a reference dataset intended for broad reuse, the authors should consider adding or releasing supplementary metadata, such as reference image date or date range, country code, confidence level, geographic coordinates, etc.

Response:

The attribute table was designed to provide the minimum required set of information. Each sample unit could be verified using different image acquisition data and this type of information was not recorded during dataset preparation. Moreover, in many cases more than one reference was required for correct class identification. In the most difficult cases, multi-year data was used to facilitate class discrimination for 2017. We assume that sample units verified by two independent interpreters provide the highest possible accuracy. The whole dataset is published as geospatial database, which allows for easy and automatic assignment of coordinates values to the sample units if needed. Country code is in fact the type of metadata which could allow for quick data filtering. This information will be added to updated dataset version.

Comment 3:

The comparison with current datasets. The manuscript briefly discusses limitations of existing reference datasets, but the comparison is general. Since the main contribution is the dataset rather than a new methodological framework, the manuscript should provide a more explicit comparison with existing European or global reference datasets (Lesiv et al., 2025, <https://doi.org/10.5194/essd-17-6149-2025>).

Response:

We agree and have added a dedicated comparison subsection and table.

Minor comments:

Line 12. "In this first step, validation sites (Sentinel-2 tiles)" The first step is to select the tiles instead of the site samples. Please clarify this terminology through the manuscript (Line 42,..). The manuscript alternates among "validation sites," "samples," "sample units," "validation samples," "tiles," and "granules." These terms should be defined once and used consistently. primary sampling unit is Sentinel-2 tile. secondary sampling unit is at 10 m pixel;

Response:

Corrected. The first stage is now described as selection of Sentinel-2 tiles as PSUs; sample units are selected only at the second stage.

Lines 17. Please add the URL of dataset at the end of abstract easy downloading.

Response:

The URL is provided in dedicated section (3 Data Records), following the journal instructions.

Line 46. What is the EO data ?

Response:

Earth Observations. The explanation is present in #46

Lines 87. Typo-fromthe. Same issue in Figure 2 caption – validationssites.

Response:

Corrected

Lines 93-94: Błąd! Nie można odnaleźć źródła odwołania.Błąd! Nie można odnaleźć źródła odwołania.Błąd! Nie można odnaleźć źródła odwołania. Correct this.

Response:

Corrected

Table1. shows the selected tiles instead of sites. Clarify this. Also define the “base” site.

Response:

“Base PSU” is now defined as a Sentinel-2 tile selected under the standard 10% per-country rule. “Added PSU” denotes the additional random tile selected for a small country when the default 10% allocation was close to zero.

Line 123. Figure 4, Add the legend showing the land cover class of each color.

Response:

The density of sample units make it difficult to distinguish which sample unit belongs to which CLC class. The aim of this figure was to present that class “other” covers big amount of area and other 13 S2GLC classes are minority in many cases.

In Table 3. Does “validation sites” mean Sentinel-2 tiles?

Response:

Yes. The terminology was corrected and unified.