The manuscript presents a dataset useful to train machine learning models for LULC mapping and the methods and results it presents are original and sufficiently well explained. However the manuscript would benefit from a clearer focus on the scope and limitations of this training dataset. For instance, both the title and the abstract do not contain 'training' and do not obviously point to the type of data that are presented. Some of the technical choices should be better justified (e.g. why not using the standard Land Cover Classification System as the basis to derive consensus between different land cover and legends; why not including the LCCS-based types from the MODIS land cover; some of the LULC categories are not defined, e.g. broadleaf cropland). The manuscript should more openly discuss the limitations and weaknesses of the dataset including on the methods applied for data validation and on the potential applications of the dataset – for instance it is not immediately obvious if the dataset could support machine learning methods to detect LULC changes with complete transition matrixes. Particularly for the classes with lowest purity – defined as the combined consensus spatially and temporally across products, if would be useful to add to the discussion some insights on what classes cause confusion and reduced consensus. I believe this (ancillary information) may be included as part of classification efforts. This information could be enriched by I recommend the publication of the manuscript after major revision – see my specific comments and suggestions below.

Title and abstract: I suggest simplifying the title and adding to the abstract more explicit reference to the actual nature and scope of the dataset. Lines 2-3: It is not entirely true that deep learning networks are unexplored for global mapping efforts (e.g. GHSL built-up areas). In this context, I believe it is more pertinent to focus on the needs for good quality training datasets in all machine learning methods. Incidentally, are there specific reason for not using the JRC GHSL- built up areas in the analysis.

Introduction: Line 21 — I suggest including here the definitions of land cover and land use, currently in lines 121-124, also adding appropriate references. Line 26 – unclear what are the biophysical properties of the land use categories. Line 28 – land cover is known as an essential climate (climate missing in the text) variable. You might want to use a synonym instead if climate variable is not appropriate. The sentence is however unclear particularly in the use of 'planet boundary'. I suggest rephrasing. In Table 1, GEE is used in the table but was not defined earlier. In Table 1, JRC Yearly History is more correctly JRC Yearly Water History. More in general, these more technical discussions (detailed reasons for discrepancies between global land cover data; Table 1 and Table 2) should be better placed in Methods (e.g. section 2.1.1). This section should instead provide the general context and present the main objectives of the manuscript and the type of data that is presented.

Methods: Figure 1 should be discussed before it is presented.

It is not immediately clear what are the concepts that guide the hierarchical system for the presentation of the LULC classes. The approach seems following the FAO Land Cover Classification System (LCCS) hierarchical approach, but LCCS is never explicitly referred to. This is important because LCCS represents the standard to harmonize land cover legends and it is used in several of the products used in the analysis (e.g. Copernicus land cover; GlobCover) so it appears strange it was not applied in the harmonization of the legends. Also, MODIS LC contains three LCCS-based types (land cover; land use and hydrology) which represent the reference land cover types for this land cover product as defined by its data producers (Sulla-Menashe et al. 2019). It seems strange that they were not included in the analysis.

The LULC categories are not clearly defined. It is especially confusing the separation of cropland classes into cereals; broadleaf and flooded. It is unclear what broadleaf cropland contains (is this permanent/woody crops?). None of the global products used in the dataset contains information on crop type, so it is not immediately clear how cereals were identified. Flooded cropland is not conceptually at the same level as cereals and broadleaf. Likewise, none of the products included the category Mangroves and there is no

indication on the distinction between Swamps and Marshlands. Thresholds for separating open/close coverage are not defined.

Line 134 (caption Table 3): I suggest changing “The numbers from 0 to 220 correspond to the class label in GEE” to “The numbers from 0 to 220 correspond to class values in the original LULC products”. A supplementary table describing the characteristics (class values; type of legend; main scope) of the 15 datasets would facilitate the understanding of the rules presented in Table 3. For each dataset, please report the appropriate link in the Data Catalog of the Earth Engine and proper citation when applicable. in the Some of the products contain both discrete and continuous categorization (e.g. Copernicus land cover) with the proportion of land cover classes in the pixel. This might be worth mentioning.

Line 135 – For products that contain only one image (P1 to P7). It is the other way round based on numbering of products in Table 1 and later in Figure 2. Line 142 – Croplands may be hardly defined as a land cover class with high temporal stability. In general, the choice of the best operator for temporal combination could be validated with some sensitivity analysis. Some thoughts should be given in the discussion as to the possible consequences for data quality and applications in the choice of the operator. Line 151 – change ‘where applied’ to were applied. Lines 151-152 – it is unclear what was done for these 5 classes and what are these classes. Section 2.1.4 and line 156 – resampling seems more accurate terms for this type of spatial operation. I suggest presenting firstly the global results with all the classes (Figure 3) and then the example. The discussion on the method applied to define the purity of the pixel could be improved and better clarified. Currently, the discussion in Lines 160-161 suggests a different meaning of purity than elsewhere in the manuscript. In here, the text refers more to the thematic classification and seems hinting to typical concepts in land cover mapping of pure *vs* mosaic land cover classes whereas elsewhere it explicitly defines purity as the spatial and temporal agreement between the various datasets.

In Figure 3, please use the full name of the classes or refer to Table 6 as you have done elsewhere. Also, it would be useful to add the labels used in the text (C1 to C28), for instance C1 – Barren lands.

Section 2.2.1 – Change ‘MODIS sensor is known by..’ to ‘MODIS sensor has high temporal coverage, ensured by Terra and Aqua satellites revisit frequencies, and also spectral and spatial features that are highly suitable for LULC mapping and change detection..’ Also, please provide a supplementary table describing the 7 bands (wavelengths). Line 184 – please explain better the reasons for not using the water flag for Permanent Snow and Cropland. Line 186 – change ‘missing-value gaps’ to ‘missing values’. Line 188 – why the maximum? Is this the highest spectral value for each band? Are there implications associated with this choice?

Line 202 – add link to GAUL dataset in GEE (see also comment above). Line 206-207 – the manuscript does not really explain why the Global modification index was included. Considering how this was produced, there is high risk of multi-collinearity and it’s not immediately clear what advantages it brings. It is mentioned later that is gives proves of the good quality of the definition of Built-up areas but this is rather vague. Line 212 – Link in footnote 2 is not working. Footnotes should be better avoided.

Line 231 – I suggest changing ‘And, the last 223 columns contain the 223 monthly observations of the time series for one spectral band’ TO ‘, The last 223 columns contain for each point the time series with the 223 monthly surface reflectance. All these values are reported separately for each of the seven spectral bands.

Line 240 – I wonder if 100 pixels are enough to assess the quality of annotation in classes that are less represented. What did guide the choice of this number? It is likely the technical feasibility and availability of resources. Please explain. What are the implications in terms of conclusions on the quality of the dataset?

Lines 241-242 – the applied thresholds may not be appropriate for some land use categories. For instance, in those pixels that contain large proportions of fallow land together with cultivated fields, these thresholds may fail to capture the dominant land use of the pixel.

Line 244 – Please provide reference to the F1 metrics. Please indicate what are the advantages and drawbacks of F1 and if alternative methods exist.

Results and Discussion – as discussed above, the limitations of the dataset for global LULC mapping and implications of the various technical choices for potential applications should be more clearly reported in a separate section. For each of the classes with lowest purity – defined in the manuscript as the combined consensus spatially and temporally across products, it would be useful to add to the discussion some insights on what classes exactly cause confusion and reduced consensus. I believe this (ancillary information) may be included in the metadata and support potential classification efforts.

Figure 7 – what variables affect the density of the time series? Considering removing this figure or explain better its usefulness.

Line 246 – report on F1 metrics seems differing to what is reported in the conclusions. Here you wrote 'As it can be noticed, as we go up from level L0 to L5 the obtained dataset accuracy increases from 87% to 96% due mainly to the forests classification'. In lines 298-299 you wrote instead 'The overall accuracy (F1 value) of the annotation varied from 96% at the coarser classification level to 87% at the finest level.'

Table 8 – I am not clear how to interpret the standard deviations that are reported for purity. Please explain.