Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-331-RC2, 2023

Comment on essd-2022-331

Anonymous Referee #2

Referee comment on "Not just crop or forest: building an integrated land cover map for agricultural and natural areas" by Melanie Kammerer et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-331-RC2, 2023

Review of Kammerer et al.: "Not just crop or forest: building an integrated land cover map for agricultural and natural areas" (essd-2022-331)

This manuscript describes a data integration effort, combining two existing vegetation/cropland cover datasets for the conterminous United States, namely the LANDFIRE National Vegetation Classification (NVC) and USDA-NASS Cropland Data Layer (CDL). As I am not an expert in the application domain for these datasets, I will focus on the integration process itself and on the validation of the resulting data product.

I think that the concept underlying this effort ("generating new knowledge by integrating existing datasets") is very valid and timely – this is how we need to deal with the myriad of geospatial data out there in order to get the maximum value out of it. The paper is well written, and from my perspective, such an effort is of potential interest for the readers of ESSD, however, there are several concerns that should be addressed prior to publication. My biggest concern is the temporal mismatch of the data (i.e., vegetation data from 2016 is integrated with agricultural data at annual temporal resolution from 2012-2020) – why not using all available Landfire epochs, or constraining the dataset to 2016 only? This temporal mismatch should be addressed in the revised version and the rationale for this decision should be clarified.

Specific comments:

- The benefit / underlying motivation of this data integration effort needs to be clarified a little bit more. While the introduction refers to some examples in the literature that assess processes playing out at the interface of cropland and natural vegetation, the Authors should add a paragraph (maybe in a concluding section) to illustrate some technical examples how these data could be used to answer specific questions – for example, one could apply convolving focal windows to identify regions where specific crop / vegetation types co-occur within a given distance. Something like this would make the contribution/value of the integrated dataset clearer.
- Also, a vectorized version (polygons) of the integrated dataset could be very useful to assess topological relationships (e.g. adjacency) between different crop / vegetation types). à If feasible, the Authors could provide such a vectorized polygonal dataset to complement their data. This will also enhance the usage of the dataset, as some researchers may prefer to work with vector rather than raster data, for topological analyses, but may not have the resources to vectorize the data.

- The validation section should be expanded. While some sort of validation has been done, little information is given about the reference data used – please provide some information (maybe a map) on the sampling locations, data source of the reference data etc. –
- Only after one hour of reviewing this paper I realized the temporal mismatch in the data, when reading this sentence "Pixels of national vegetation are the same in all rasters provided here and represent land use in 2016." on the data website (https://data.nal.usda.gov/dataset/data-not-just-crop-or-forest-building-integrated-land-cover-map-agricultural-and-natural-areas-spatial-files/resource/8c92879b-92cf-4e86-a3c4-0e672007a1df) - NVC data from 2016 is integrated with CDL data annually from 2012-2020? This is not clear from the manuscript. What are the implications of keeping vegetation cover stationary over time? Does cropland change faster than vegetation? How does this temporal mismatch affect the usability of the integrated dataset? Can it be used to assess recent processes at all? This issue needs to be highlighted and thoroughly discussed. When looking at this page: https://landfire.gov/data_overviews.php, I see that Landfire has been released in several years besides 2016 – why did you not integrate Landfire and CDL in annual pairs, for the years available? Please provide a rationale for this. This is probably my biggest concern about this manuscript.
- Same sentence on the data website : "Pixels of national vegetation are the same in all rasters provided here and represent land use in 2016." --> isn't vegetation land cover, instead of land use?
- If not done already, Authors should provide a spatial layer (raster dataset) of the mismatched / unresolved pixels, so that users can include these discrepant areas explicitly in their analyses.
- Related to that, it is not clear to me in which year the agreement assessment was conducted (sorry if I missed it), and whether the stats (e.g. 5% of conflicting pixels) refers to a specific year, is it 2016?
- It seems that the union of agricultural land use and vegetation land cover is used as the analytical "universe" in this study. How do these areas relate to other land cover / land use types, such as urban areas / developed land? It would be great if the Authors could conduct a cross-comparison to a "spatially exhaustive" dataset such as the NLCD – how does the integrated dataset agree with the classes from NLCD? This would be some kind of "external" evaluation, while the agreement analysis would be an "internal" validation. Perhaps a cross-tabulation of the area proportions per crop/vegetation class and NLCD land cover class would be interesting (see e.g. Fig 4 in https://www.nature.com/articles/s41597-022-01591-0)
- I suggest to rename the dataset containing the uncertainty statistics. Please change "tabular data" to "uncertainty statistics" or similar.
- Lastly, I suggest to come up with a name for the integrated dataset. This will make it easier to refer to the dataset and ultimately, increase the visibility of the product.

Minor comments:

Sorry if I missed it, but what is the spatial resolution of the input data? I think it is 30m for both of the datasets, this should be stated in your geoprocessing section. Also, it is important to know whether the raster grids align, or is there an offset between the two grids? Did you have to resample one of the layers to the grid of the other layer? If so, how was this done (nearest neighbor resampling?).

Some terminology… AFAIK one would speak of "forest land cover" on the one hand, but "agricultural land use" on the other hand – in the title and in the manuscript you write "land cover" – the term "land use" does not occur in the manuscript. However, isn't your integrated dataset truly a LULC (land use / land cover) dataset? I think the integration of land cover and land use should be highlighted more in the paper (and maybe even in the title).

Minor detail: Some of the maps in Figures 2,3,4 show areas / area proportions and thus, should use an equal-area projection rather than showing Lat-Lon in a cartesian coordinate system. Lat-Lon are angular coordinates and should IMHO not be shown in cartesian coordinate systems, in particular when it comes to mapping areas / densities / area proportions – as a pixel in Maine has a different area than a pixel in Florida. I suggest to use Albers Equal Area projection.

Fig. 4- top map: I don't think it necessary to show an "empty" map?

The introduction should include a short synopsis of similar data integration efforts, to illustrate that this is a trending topic in general, and across disciplines. For example, a similar effort from the field of human settlement modelling would be this: https://www.tandfonline.com/doi/full/10.1080/17538947.2018.1550121