Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-135-RC2, 2022

Comment on essd-2022-135

Anonymous Referee #2

Referee comment on "Four-century history of land transformation by humans in the United States: 1630–2020" by Xiaoyong Li et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-135-RC2, 2022

The manuscript reconstructed land use and land cover history during 1630-2020 over the conterminous United States (CONUS). It gathered multiple sources of data including remote sensing-based land cover maps, national inventories and statistics data, meteorological fields, topographical data and others. The resulting dataset has an improved spatial resolution than other global datasets, potentially facilitating regional modeling work. However, the manuscript is not well organized. For example, in the method section, there are missing details about how various sources of input data were combined and adjusted to generate full time series of urban and crop land area. And also, the validation needs more justification on the choice of spatial scale and more discussion about what causes differences between the data from this work and other datasets. For example, this data shows a different trend in areas of urban and pasture in the past two decades compared to others. I have provided detailed comments below.

P2, L55, LUH2 has a spatial resolution of 0.25-deg.

P2, l55, please provide references to the ‘substantial uncertainties’ statement.

P5, L95, it would be clearer to split the datasets by the purposes in your model (e.g., input vs validation). And also, there is no temporal resolution for each dataset. Apparently, some of them are not annual.

P6, L100, please clarify if all land cover types of developed land (i.e., Developed, Open Space; Developed, Low Intensity; Developed, Medium Intensity; Developed High Intensity) were regarded as urban land.

P6, L100, did you take an average of urban land area during 2001-2016 and use it as the baseline? Or you did make use of the time series of urban land areas. I wonder about the temporal variation of urban land per capita and its impacts on estimation of historical urban areas. It would be better to include this information in supplements.

P6, L106-108, which dataset were these criteria applied to? CPHR, CAHA, or other?

P6, L112, I could not find the CPHR dataset in Table 1. Please make sure the product name and time period is consistent between the table and manuscript.

P6, L117, 1975-2020, right? If not, please list this dataset clearly in Table 1. Please add more details about how the adjustment was done to make 'the inter-annual variation more reasonable'. For example, was this adjustment done at national scale then disaggregated to state scale?

P6, L125-126, what causes this difference? Due to the CAHA crop harvested area?

P6, L127, please justify why the cropland per capita was calculated at national level instead of state level like the urban area per capita. This could avoid any potential confusions about the following assumption you made.

P7, L130-134, was there a harmonization process applied to connect cropland area during 1630-1880 derived from the HYDE and cropland area during 1889-2020 derived from the CAHA? If not, please add time-series plots of derived cropland areas that show such harmonization is not needed.

P7, L141-L143, please clarify how the pasture per capita during 1630–1982 was estimated from pasture per capita at the state level (NRI data-based) in 1982 and the national level (HYDE data-based).

P7, L157, "... then multiplied the forest area from USDA-FR to generate …". What is the time period of USDA-FR forest area used here? 1630?

P8, L170, why is there no scaling factor for TA_t_r (s) < TLA(s) ? How was the difference/residual between TA_t_r (s) and TLA(s) dealt in your following analysis? And also, what is the source of the state's total land area (TLA)?

P9, L182-184, please add references to how ANN can be used to solve the nonlinear geographical problems.

P9, L184-185, please add more details about how land use probability was generated from ANN and NLCD. And also, NLCD is supposed to be an independent variable used in the modeling, right? However, NLCD is a land cover data, how can it be useful for modeling of land use probability? Please keep in mind that land use and land cover are usually used interchangeably but are actually different terms. If you regard them as the same terms, why do you refer to it as land use probability instead of land use and land cover probability?

P9, L200, what are the difference between ES_weight_t and SE_weight_t in Eq. 4 and 6.

P9, L200, what are the values of t1 and t0 in Eq. 7? Are they the starting and ending year of each subperiod?

P10, L216, what does 'Boolean type' mean? Categorical type like NLCD that each grid (i.e, 30 m) has a single land use and land cover type?

P10, L217-218, please elaborate more on "the total number of potential pixels or the land use demand was determined based on the reconstruction results in Section 2.2. Then, the area difference of land use type k between the target and current map was calculated." The resulting data from section 2.2 is the state level total area of urban, crop, pasture and forest, right? So, such 'area difference' is at the state level, right?

P10, L228-236, the LUH2 used for validation, but it was not mentioned in section 2.1 and Table 1.

P10 L229, please justify why the validation against NLCD was at state-level. As you highlighted that your land use data is at 1 km, a finer spatial resolution than most other data, so the validation at 1 km will be more informative about the value of this dataset. A good agreement on state-level total land use area does not necessarily indicate the spatial allocation of total area to 1 km is good as well.

P11, Figure 4. Such comparisons to different datasets are very important and informative. However, I would suggest more discussion on the 'outlier' states in figure 4a and 4b. For example, please dig into which states your estimates of urban land area is lower than HISDAC? And what are the potential causes?

P14, L290-L300, is the transition in figure 7 the gross transition or net transition? Please clarify this somewhere because their magnitude and impacts on climate and carbon modeling are quite different. For example, 30% deforestation and 30% reforestation have zero transition on forest, but the biophysical effects could not be ignored.

P17, Figure 9d, the black line stops at near 1920.

P17, Figure 9a and 9c, even though NLCD is the input of urban land area to your method, there is a difference in temporal trend in urban area between NLCD and your dataset. Please discuss more about what assumptions cause such trend difference, and how it affects the reconstruction before the 2000s. Same discussion is needed for the pasture as the trend is even opposite.

P20, L382-383, please explain why inventory-based data is more reliable than the satellite-based forest (NLCD) and biomass density-based forest (LUH2). Forests could have different definitions by various sources, and forest areas are subject to the definition, thus I could see which definition/data source is more reliable than others.