Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-256-RC1, 2022


# Comment on essd-2022-256

Anonymous Referee #1

Referee comment on "Location, biophysical and agronomic parameters for croplands in northern Ghana" by Jose Luis Gómez-Dans et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-256-RC1, 2022

Summary: In the study titled “Location, biophysical and agronomic parameters for croplands in Northern Ghana.”, the authors combine in situ observations from two large measurements campaigns in Ghana with Earth observation data to generate a dataset of crop location, biophysical parameters (including LAI and leaf chlorophyll concentration), crop yield, and biomass over smallholder maize farms. I appreciate the huge amount of work done by the authors and think the dataset can be useful for future science and applications. In particularly, getting field-specific data for these parameters can be relevant for training and validation of larger-scale products and also for continuous monitoring of these parameters over time to optimize farming practices in this region. However, I am concerned about several sources of uncertainties in the dataset, which depend on both the data sources chosen and the methods used. See my comments below.

**Major Comments:**

- One major source of uncertainty is the selection of data products. For instance, the authors choose PlanetScope surface reflectance data to derive the vegetation indices, which is then used to estimate LAI. They mention that there is a lot of noise in this dataset, which requires a couple of rounds of outlier detection and removal and that the positional accuracy can be as large as 10 m. If so, why not used Sentinel surface reflectance to estimate NDVI? Sentinel has a regular acquisition time, lower positional errors, and more consistency of atmospheric corrections. The final dataset is at around 10 m resolution, so I am unsure about the added benefit of Planet imagery here.
- Another source of uncertainty related to data products is the landcover mask used based on the ESRI global 10 m land cover dataset. Why was this classification dataset chosen instead of other similar 10 m land cover datasets (Venter et al. 2022)? Was the same ESRI landcover classification used for both years or did the authors use the 2020 and 2021 land cover products (https://planetarycomputer.microsoft.com/dataset/io-lulc-9-class) separately? I am concerned how using these difference datasets would impact the final results and datasets produced.

- The second source of uncertainty is regarding the methods used and how they are impacting biases in the final dataset. One concern is about the outlier detection. This is done in a more or less statistical manner. However, is there a way to check with the in situ observations whether the outliers are a real signals or noise. Here, it would also be good to see the vegetation index from Sentinel in Fig. 5. If the outlier is purely due to the uncertainties in the PlanetScope estimates, it might be better to use Sentinel for calculating the NDVI?
- The overall accuracy of the derived LAI and the NDVI to LAI are both quite low (Fig. 9) with a correlation coefficient of 0.49 (so r2 of around 0.25). Is this a reasonable accuracy for such a dataset and how would end users justify using this dataset if such a low proportion of the variance is being explained? Here, I am also surprised why the authors showed the r value in Fig. 9 and the r2 value in Fig. 14. Best to be consistent.
- Looking at Fig. 10, there are both systematic biases and differences in phenology between predicted and field LAI. Is this bias somehow incorporated in the final dataset? It would be helpful to include some indication of this bias so that end users know what the uncertainties are over a field before they use the results.
- How do this dataset addresses the issue initially raised in the introduction. As an example, in the introduction, the authors talk about the limitations of remote sensing due to the presence of trees, inter-cropping practices, etc. But then they choose the fields with the least amount of tree cover and inter-cropping for the in situ crop measurements. I think the authors need to expand upon this discussion or modify the introduction.

**Minor comments:**

- It is unclear how comprehensive this dataset is. What is the fraction of the total area of the smallholder maize fields in Ghana that this dataset pertains to?
- Line 115: Here and elsewhere, probably best to be explicit that these are in Celsius.
- Line 200: saturation effects with high what?
- Line 230: How were the pixels split? Randomly? Some kind of stratification? Was there only one set of training/validation? Why not use multiple random splits to check for consistency of results?

**References:**


- Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T., & Singh, G. (2022). Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover. Remote Sensing, 14(16), 4101.