Interactive comment on "Development of a standard database of reference sites for validating global burned area products" *by* Magí Franquesa et al.

Anonymous Referee #3

Received and published: 27 August 2020

General Comments

This manuscript describes a first attempt at compiling a common database of burned area reference perimeters ("BARD") suitable for validating remotely sensed burned area data sets. While the goal of producing the BARD is laudable, I feel the result falls somewhat short in that the authors provide no guidance in how this dataset should be used in practice. While reprojected and vectorized into a common format, the six underlying validation data sets were not generated in an entirely consistent manner and vary significantly in terms of sampling strategy and minimum mapping unit and

various other important respects. As a result, I feel it is essential for the authors to advise users how the database as a whole should be used. For example, should some or all of the individual data sets be merged, or should they always be used separately? If the latter, then any validation of a global data set will yield six different sets of validation results. How should these results be interpreted, especially for the individual data sets that overlap in space and time, such as FireCCI Africa and FireCCI Africa S2? Furthermore, the authors state that "The Burned Area Reference Database will be expanded with new reference files that are being produced in the FireCCI project and we encourage future contributions from the scientific community", but it is not clear how this plan can scale practically as the number of data sets grows.

Specific Comments

L41: "they require generating global reference data that is based on higher-resolution sensors" Although I agree with this statement, it overlooks studies such as Roteta et al. (2019) who used 30-m Landsat images to validate a 20-m Sentinel-2 burned area product.

L59: Giglio et al. (2018) give a release date of mid 2008 for the NASA MCD45A1 product.

L68: "The MCD64A1 Collection 5 was not formally validated" Giglio et al. (2009) performed an "accuracy assessment for three geographically diverse regions (central Siberia, the western United States, and southern Africa)" using 50 Landsat scenes. Is this not validation?

L85/Section 2.1: The authors note the importance of sampling design and describe

various important components of this process, but not all of the BARD data sets seem to have adopted the strategies described in this section. It would be helpful to note any deviations within BARD from the sampling strategy described here. The authors might perhaps also provide a brief summary of the CEOS-LPVS validation stages to help readers interpret the stage numbers mentioned later for the individual data sets (in Table 2, for example) in the context of sampling.

L158: "The FireCCI global 2008 dataset includes 129 reference data files" This number differs slightly from Padilla et al. (2014), who refer to "102 sampled pairs". Presumably additional scenes were added to that data set. This is worth mentioning since it would alert readers that the summaries and/or statistics provided in Padilla et al. (2014) do not necessarily apply to the FireCCI global (2008) distributed in BARD.

L195/Section 2.4.4: The 2016 FireCCI Africa S2 data set is not mentioned in either of the references cited in this section. Please add the correct reference or clarify that the data set has not been previously published.

L208/Section 2.4.5: Rodrigues et al. (2019) mention a minimum mapping unit of 21 ha. Does this threshold also apply to the BrFLAS Brazil data distributed in the BARD?

L230/Section 2.4.6: Hawbaker et al. (2020) include the following remark about the BAECV validation data set: "Because no independent reference data were available for burned areas in agricultural cover types, the Landsat-based BAMS reference dataset did not train on agricultural fires and consequently cannot be considered accurate for this cover type." Have the unreliable reference polygons belonging to this category been flagged or removed from BARD? If not, some guidance to users about how they should identify and handle such cases would be appropriate.

L242: "The pre- and post-fire image pairs did not specifically represent a probability sample within a year but were designed to target changes incurred over the peak fire season." Given this targeting of the peak fire season, is it appropriate to use this data set for assessing out of season commission errors?

L268/Section 2.4.7: Given that the NOFFi-OBAM mapping service "is activated after large wildfires events and under explicit requests by the local forest offices", is it appropriate to use this data set for assessing commission errors? Please explain and include appropriate caveats if necessary.

Figure 3: This figure shows perhaps a dozen validation sites that are not shown in the equivalent figure of Padilla et al. (2014), where the 2008 FireCCI global validation data set was originally described. Please see related L158 comment above.

Figure 5 would be much more useful if it included clouds or some other source of missing data in the Landsat image stack. The long unit sampling is not clearly described in the manuscript, but I think I understand most of what the authors here poorly describe only after consulting Figure 12 of Padilla et al. (2018). Perhaps the authors could include a similar figure here.

Figure 9: Not clear why it is useful to highlight FireCCI TSAs vs. Added TSAs on the map. It would be more useful and more consistent to show the time period between Landsat image pairs as was done for the other data sets in Figures 3, 6, 7, and 8.

Table 2: Please show the total areas of the separate burned, unburned, and no-data

classes for each data set.

BARD DOI landing page (https://edatos.consorciomadrono.es/dataset.xhtml?persistentId=do
The landing page describes BARD almost exclusively as a FireCCI effort. This is a
little bit inconsistent with the manuscript, which says that the database "was created by
compiling existing reference burned area datasets from different international projects."

Technical Corrections

L40: change "sensors" to "sensor"

L57-58: Acronyms MERIS and MODIS not defined

L85: change "amount" to "number"

L91: change "sample" to "samples"

L213: change "covering the 77%" to "covering 77%"

Figure 2 caption: change "Time distance between" to "Time period between"