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.

Response: We have now included some reflections and information (Tables A1-A4) on practical uses of the database. For further details, the reader is also referred to the articles where each dataset was first published. We consider this as a collection of BA reference datasets, not as a single one. Therefore, is up to the user to select certain regions or periods to produce his-her accuracy estimates. The uncertainty of accuracy estimates should contextualize the discrepancies between validation results from several datasets (and same product and overlaps in time and space). Slight discrepancies are expected as any single dataset is observing a sample of reference data instead of the whole population. We have now provided additional data to compute those accuracy metrics, but this database can be used in several different ways by potential users. Some, for instance, may use certain datasets for training their algorithm and some others for validation. Obviously, we do not aim to convert the BARD in a standard validation source, but just to provide useful data for BA algorithm developers and modellers.

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.
Response: The Roteta et al. paper performed a stratified random selection of Landsat images for generating the reference perimeters to compare accuracy metrics of S-2 and MODIS BA products. A previous validation based on a systematic sample of S-2 MSI images gave similar results, so only the last validation was included in the paper. It is certainly more convenient to use higher resolution images for validation, but in this case it was decided to use the same validation dataset to make comparisons between coarse and medium resolution sensor products more fair. In addition, a statistically design sample based on high-resolution images (Planet) is very complex and costly, and when using them for BA validation have been done in a very qualitative way (Roy, D.P., Huang, H., Boschetti, L., Giglio, L., Yan, L., Zhang, H.H., & Li, Z. (2019). Landsat-8 and Sentinel-2 burned area mapping - A combined sensor multi-temporal change detection approach. Remote Sensing of Environment, 231, 111254.)

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

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?

Response: Giglio et al. (2009) selected three different areas to represent different ecological conditions to evaluate their algorithm and no probability design was applied. The authors provided only the producer's accuracy for the scenes previously selected but didn't report global accuracy estimates of the product.

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.
Response: Section 2.1 aims to provide a general overview of the sampling design methodologies developed for burned area validation. The particular sampling design adopted for each dataset is specified in the correspondent description of the datasets in section 2.4 and summarized in table 2.
Thank you for the suggestion, we have provided a description of the CEOS-LPVS validation stages.

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.
Response: The sampled units of such dataset comprises 105 units and the correct reference for this dataset is Padilla et al. (2014, 2015), the rest of the reference files (24) shouldn't be included in the dataset. The dataset has been updated including only these 105 reference files, and the dataset description has been updated accordingly.

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.
Response: This dataset was used to perform and initial validation of the FireCCISFD11 product but has not been published. We have indicated this situation in Table 2 where we provide the related publication of each dataset.

L208/Section 2.4.5: Rodrigues et al. (2019) mention a minimum mapping unit of 21ha. Does this threshold also apply to the BrFLAS Brazil data distributed in the BARD?
Response: No, no minimum mapping unit was applied to the BrFLAS Brazil. In any case, this dataset has been removed from the BARD since it does not follow CEOS cal-val standards. Please see short comment 6 and response.

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.
Response: The CONUS Landsat Burned Area (previously named BAECV) reference dataset classifies agriculture cover types as burned/unburned. The comment in Hawbaker et al. (2020) was made to acknowledge that because we lacked ancillary datasets in agriculture areas, the reference dataset burn classifications were not explicitly trained using agricultural burned polygons, and therefore, the reference dataset may be less accurate in this cover type. As 19 of the 28 TSAs contain at least some agricultural area it does not make sense to remove these shapefiles, however, in response to this comment we have added a sentence in the description of this dataset of the reviewed manuscript:

...The low-, medium- and high-intensity development classes (i.e. urban areas) were masked out using the National Land Cover Database (NLCD, https://www.mrlc.gov/national-land-cover-database-nlcd-2016) (Homer et al., 2015) to reduce spectral confusion between burned areas and impervious surfaces. *Similarly, agricultural burns were not used to train the reference data burn classification, therefore the accuracy of the reference dataset in agricultural areas is unknown. If this is of concern to users, then users can mask the "cultivated crops" land cover type from the reference data using the NLCD* "

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 dataset for assessing out of season commission errors?
Response: According to FireCCI51, the main peak fire season for CONUS goes from July to September-October. 80.36% of reference files from CONUS Landsat Burned Area dataset include months out of the fire season. Thus, we consider that this dataset is appropriate to assess Ce out of fire season.

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.
Response: Yes, NOFFi-OBAM is appropriate for assessing commission errors as reference data follow CEOS cal-val standards. As we explain in the dataset description: 'The NOFFi-OBAM fire perimeters were used as basis for creating the reference data for the NOFFi Greece reference dataset' and we mention that 'Small fires within the specific time series that were not mapped from the NOFFi-OBAM service were explicitly digitized'. Additionally, unburned and unobserved categories were added to adapt this product to the CEOS cal-val standards.

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 dataset was originally described. Please see related L158 comment above.
- The figure has already been corrected according to L158 response.

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.
Response: Thank you, we have modified figure 5 to clarify the schematic process to obtain long unit reference data. We also have extended the explanation on how long units are obtained in section 2.2.

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 be-tween Landsat image pairs as was done for the other data sets in Figures 3, 6, 7, and 8.
Response: The CONUS Landsat Burned Area dataset used 28 validation sites that were repeatedly sampled in each of the six validation years (with different time gaps in each year), making it challenging to provide a figure similar to Figures 3, 6, 7, and 8. This is the only dataset that was created to validate a specific region (CONUS) based on a previous existing global dataset (FireCCI global 2008) and this is a relevant aspect we mention in the reviewed manuscript:
'*another key advantage of stratified random sampling design that should be strongly emphasized is that it makes it possible to increase the sample size of an initial global sample for specific regions or rare land-cover classes (Stehman et al., 2012). This is the case of CONUS Landsat Burned Area (1988-2013) dataset where reference sites for the CONUS extent were augmented*

based on the initial sample of the FireCCI global (2008) dataset.'. Figure 9 emphasize this property. In response to this comment we have added additional text to the Figure caption to clarify.

"Reference data were generated for each TSA in each of the six sample years (1988, 1993, 1998, 2003, 2008, 2013)."

Table 2: Please show the total areas of the separate burned, unburned, and no-data classes for each data set.

Response: We have added this information in a separate table (table 3) as suggested in SC1.

BARD DOI landing page (https://edatos.consorciomadrono.es/dataset.xhtml?persistentId=doi:10.21950/BBQQU7). 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."

Response: Yes, BARD is an initiative that arises from the FireCCI project and 92% of reference files were produced in the FireCCI project. However, we consider essential the present and future contributions of other initiatives to this effort.

Technical Corrections

L40: change "sensors" to "sensor"

Response: Done

L57-58: Acronyms MERIS and MODIS not defined

Response: Done

L85: change "amount" to "number"

Response: Done

L91: change "sample" to "samples"

Response: Done

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

Response: BrFLAS dataset has been removed from the database since it does not follow CEOS cal-val standards. Please, see short comment 6 and response.

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

Response: Done