General

The reference datasets for validating global burned area products provide a valuable resource to the fire mapping community. As the authors note, collecting reference data to validate burned area products is an expensive and time consuming proposition. Having available a vetted set of reference sample sites for map producers to readily access will greatly enhance the quantity and quality of information available to assess and compare accuracy of burned area products. The global extent of these datasets will facilitate regional comparisons as well, as users of the data will be able to extract data specific to their study area. One of the fundamental challenges of mapping of any theme, burned area or otherwise, is the immense difficulty of obtaining reference data. The burned area reference dataset (BARD) presented by the authors is a significant advance to diminish this difficulty.

We sincerely appreciate this review and thank the positive comments about the contribution of this manuscript to the field.

Specific Comments

1. The authors astutely identify the role of sampling in the collection of these burned area reference datasets (Line 63). It would be useful to add some explanation distinguishing between reference data collected by a formal sampling design, often called probability sampling designs, and reference data collected by convenience, ease of access, or other method that does not necessarily have randomization. Reference data collected by a randomized sampling design are suitable to support rigorous statistical statements about accuracy, whereas data collected by convenience can be suspect in this regard (i.e., data may not be representative of the entire area of interest). The implications of how the reference data were obtained should be noted. The manuscript clearly indicates that the Boschetti et al. (2019) and Padilla et al. (2014:2015) reference datasets were obtained from locations selected by stratified random sampling. For some of the other datasets, this is less clear. It would be useful for the authors to check each dataset and be sure that it is indicated whether the dataset had an underlying randomized sampling design.

Response: This is indeed a very useful add in to the description. We have extended section 2.1 to better explain these aspects. We have revised the detailed description of each dataset and have included which sampling method was used. Please, also note that the sampling design used in each dataset has been summarized in table 2.

2. Related to the previous comment, the manuscript identifies that several of the datasets included were selected by stratified sampling designs, and these designs had intensified sampling in high burned area strata. According to the original articles associated with these datasets, rather complex estimation formulas have to be applied to such data (i.e., the less complicated formulas of simple random sampling are not appropriate when the sampling was stratified with different sampling intensities in the strata). It would therefore seem necessary that users of these reference datasets be cautioned about the need to use proper estimation formulas if users are to correctly report accuracy from these stratified sample datasets. This would also create the need to include in the datasets the information required to apply these estimation formulas, for example stratum sizes, the stratum ID of each sampled unit, and perhaps additional information depending on the specific details of the particular dataset.

Response: Thank you for this observation. We have added the required data to use the validation datasets obtained through SRS to make probabilistic estimations of accuracy. The stratum ID of each sampled unit and the total area of the TSAs from which reference data was obtained have been added to the .csv files provided in the metadata folder. In addition, a table with the stratum sizes for each reference dataset is also provided in section 5 (Appendix A: Supplementary tables).

Technical Corrections and Suggestions

Throughout, readability would be enhanced by using paragraph indents at the start of each paragraph.
Response: Preprint manuscript was prepared according to the journal template, post-editing will be applied to the final version.
L23: insert “a” to revise to “requires a high level”
Response: Done.
L26, L29: Given that the acronym BARD was defined at Line 26, replace “The Database” with “BARD”
Response: Done.
L40: “sensors” should be “sensor”
Response: Done.
L41: revise to “reference data that are based on” [“data” is plural so “data that are”]
Response: Done.
L46: “products” instead of “product efforts”
Response: Done.
L63, L79, L105, L106, L159, L161, L164, L198, L205, L207, L209: Throughout the manuscript, the words “file” and “files” are sometimes used to the refer to the actual reference data. For example, at L63, the “files” were not derived from pairs of images, but rather the “reference data” that are stored in the files have been produced from the pairs of images. The text should be revised to replace “files” with “reference data” unless the text is referring to the actual files that store the reference data.
Response: Thank you for this observation, we have revised the document and changed it as suggested.
L64: Replace “without probabilistic meaning” by “that were not selected using a probability sampling design”. It is not clear what “direct sampling” is. Is direct sampling convenience, purposeful, or other sampling without randomization?
Response: We have modified the sentence to ‘*Early validation exercises were subjected to a first stage validation, usually based on small samples of reference sites that were not selected using a probability sampling design, but rather by a purposeful or convenience selection based on data availability or expert knowledge to ensure diverse wildfire conditions were included in the sample*’.
For all examples at Lines 65-70, it appears that there was a rationale for why sites were selected (even if they were not selected by a randomized protocol). It would be useful to mention what purposeful selection criteria were used. The Roy and Boschetti example mentions sites selected to be spatially distributed across the landscape, so this is an example where the manuscript provides useful additional information regarding the purposeful selection criteria.
Response: Validation sites selection in Chuvieco et al. (2008) was based on Landsat and CBERS images donated by regional space agencies, when Landsat archive wasn’t free open to the public. We have mentioned it in the corresponding paragraph.
L70-71: If Boschetti et al. (2019) collected data for only a single year, does that qualify as a “full spatio-temporal validation”? It would be helpful to define what a “full validation” is in regard to time and space.
Response: We have removed the expression ‘full spatio-temporal validation’ to avoid confusion and changed the sentence to ‘*A recent study has provided a validation of the MCD64A1 product implementing a probability sampling design and using Landsat-8 Operational Land Imager (OLI) images but only for a single year (Boschetti et al., 2019)*’.
L88: insert “design” after “random sampling” to create “stratified random sampling design”
Response: Done.
L89: Consider revising to: “Boschetti et al. (2016) extended the sampling design to include the temporal dimension of the sampling units.”
Response: Done.
L90: insert “the” between “allocate sample” and delete “a” from “example a stratified”
Response: Done.

L91: insert “the” before “sample”

Response: Done.

L94: replace “are” by “is” because “dimension” is a singular noun

Response: Done.

L99: delete “a”

Response: Done.

L106: Consider revising to: “The procedures implemented to obtain those burn patches are diverse, depending…”

Response: Done, text has been modified as suggested.

L109-110: Consider revising to: “Parts of the scene that cannot be observed or interpreted because of clouds or sensor problems (i.e., Scan Line …”

Response: Done, text has been modified as suggested.

L115: replace “such” by “each” and replace “like” by “such as”

Response: Done.

L153: Are n=127 and n=131 the number of TSAs sampled? It is not clear what these numbers represent.

Response: The numbers refer to the number of images interpreted from each sensor, 127 images from Landsat-5 and 131 from Landsat-7. We have changed the sentence to clarify this point. Please, note that numbers have been modified because we initially included some reference data that shouldn’t be included in this dataset.

‘A total of 210 images from Landsat-5 TM (n=101) and Landsat-7 ETM+ (n=109) satellite sensors were used to retrieve BA perimeters’.

L170: delete “to each sample unit” because this threshold is applied to all TSAs. That is, all TSAs are assigned to strata as part of the sample selection process. It is not just the sampled units that are assigned to strata.

Response: Thank you for pointing that out, we’ve changed the text according to your observation.

L172: given that “proportional allocation” for stratified sampling is defined as the sample size in each stratum being proportional to the number of units in the entire study region belonging to that stratum, replace “applying a proportional allocation” by “applying a sample allocation”.

Response: Done.

L182: replace “in” with “of” and replace “days” with “day”

Response: Done.

L185: It is not clear how the actual time period covered by these “long units” is defined.

Response: The long sampling units are defined by multiple consecutive pairs of images (short sampling units, separated by 16 days or less) covering at least 100 days. We have clarified the concept of short and long units in section 2.2.

L186: Consider revising to: “Reference maps using long units concatenate consecutive
8-16 day maps (Fig. 5).”

Response: This line has been removed as long unit reference data generation methodology is now explained in section 2.2.

L188: The 50 units are for fire CCI Africa compared to 100 units per year for FireCCI
global?

Response: The authors used different sampling intensities for Africa and global. The 50 (long) units for FireCCI Africa implied an effort in the generation of reference data similar to that for the 12 years of FireCCI global. In the former case, 1052 pairs of images were processed, and on the latter case 1200.

L189: replace “consists on” with “consists of” and replace “perimeters” by “perimeter”

Response: Done

L190: replace “units” by “unit” (2 cases) and “days” by “day”

Response: Done

L198: remove “A” before “systematic sampling”

Response: Done

L201: replace “the whole” with “all” and replace “was” with “were”

Response: Done

L203: “consecutively” should be “sequentially”

Response: Done

L209: “joined” should be “joint” and “by” should be “between”

Response: BrFLAS dataset has been removed from BARD (please, see SC6 response)

L213: delete “the” before “77%”

Response: BrFLAS dataset has been removed from BARD (please, see SC6 response)

L219: replace “scar samples” by “scars sampled”

Response: BrFLAS dataset has been removed from BARD (please, see SC6 response)

L223: “days” should be “day”

Response: BrFLAS dataset has been removed from BARD (please, see SC6 response)

L224: “pair” should be “pairs”

Response: BrFLAS dataset has been removed from BARD (please, see SC6 response)

L228-229: Continue to use the same phrasing as at L180 and L207 to identify the stage of the reference dataset. The sentence structure at L180 and L207 is much easier to read.

Response: BrFLAS dataset has been removed from BARD (please, see SC6 response)

L231-232: replace “generated to perform the validation of the BAECV” with
“generate to validate the BAECV”

Response: As we have renamed BAECV dataset to CONUS Landsat Burned Area, the sentence “generated to perform the validation of the BAECV” has been replaced by “generate to validate the Landsat Burned Area product”

L232: Move the text “Landsat Burned Area Essential Climate Variable” to before the first use of BAECV at Line 231.

Response: As we have renamed BAECV dataset to CONUS Landsat Burned Area, the sentence has been changed to: ‘*The Landsat Burned Area reference dataset (Vanderhoof et al., 2017;2020) extends across the contiguous United States (CONUS) and was generate to validate the Landsat Burned Area product (Hawbaker et al., 2017;2020)*’.

L238: delete “A” before “systematic”

Response: Done

L239: the three values of n sum to 335 images not 336

Response: Thank you, the error has been corrected

L243: replace “…only two (pre and post-fire image…” by “…only two images (pre and post-fire) …”

Response: Done

L266-267: Continue to use the same phrasing as at L180 and L207 to identify the stage of the reference dataset.

Response: Done

L272: “wildfires” should be “wildfire”

Response: Done

L279: “were” should be “was”

Response: Done

L283: “postfire” should be “post-fire”

Response: Done

L284: “formers” should be “former”

Response: Done

L290-291: Continue to use the same phrasing as at L180 and L207 to identify the stage
of the reference dataset.

Response: Done

L306: Consider changing “futures updates come to replace the lack…” with “future updates remedy the lack…”

Response: Done, text has been modified as suggested.