##

Anonymous Referee #1
Received and published: 21 June 2020

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.
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.
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.
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”
Done.

L26, L29: Given that the acronym BARD was defined at Line 26, replace “The Database” with “BARD”
Done.

L40: “sensors” should be “sensor”
Done.

L41: revise to “reference data that are based on” [“data” is plural so “data that are”]
Done.

L46: “products” instead of “product efforts”
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.
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?
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.
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.
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"
Done.

L89: Consider revising to: "Boschetti et al. (2016) extended the sampling design to include the temporal dimension of the sampling units."
Done.

L90: insert "the" between "allocate sample" and delete "a" from "example a stratified"
Done.

L91: insert "the" before "sample"
Done.

L94: replace "are" by "is" because "dimension" is a singular noun
Done.

L99: delete "a"
Done.

L106: Consider revising to: "The procedures implemented to obtain those burn patches are diverse, depending…"
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 …"
Done, text has been modified as suggested.

L115: replace "such" by "each" and replace "like" by "such as"
Done.

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

L182: replace "in" with "of" and replace "days" with "day"
Done.

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

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

L198: remove “A” before “systematic sampling”
Done.

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

L203: “consecutively” should be “sequentially”
Done.

L209: “joined” should be “joint” and “by” should be “between”
BrFLAS dataset has been removed from BARD (please, see SC6 response).

L213: delete “the” before “77%”
BrFLAS dataset has been removed from BARD (please, see SC6 response).

L219: replace “scar samples” by “scars sampled”
BrFLAS dataset has been removed from BARD (please, see SC6 response).

L223: “days” should be “day”
BrFLAS dataset has been removed from BARD (please, see SC6 response).

L224: “pair” should be “pairs”
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.
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”

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.
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"
Done.

L239: the three values of n sum to 335 images not 336
Thank you, the error has been corrected.

L243: replace "…only two (pre and post-fire image…" by "…only two images (pre and post-fire) …"
Done.

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

L272: "wildfires" should be "wildfire"
Done.

L279: "were" should be "was"
Done.

L283: "postfire" should be "post-fire"
Done.

L284: "formers" should be "former"
Done.

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

L306: Consider changing "futures updates come to replace the lack…" with "future updates remedy the lack…"
Done, text has been modified as suggested.

##########

Anonymous Referee #2
Received and published: 4 August 2020

General Comments

This manuscript describes the availability of a new dataset comprising a compilation of reference burned area data, which can be used for the validation of burned area products. The short

description paper outlines the methods used to standardise a number of different datasets into a common format, and a more detailed description on each one. It also gives an overview of why validation is necessary but not always readily available, which provides useful context.

Validation of burned area products is definitely lacking in the field, and this is a welcome contribution to the research area. I think it will be useful for many researchers working on fire and burned area. The methods are well-referenced, and are mostly clearly described, with the exception of a few points as outlined below. The data is readily available via the link provided in the text, and can be accessed immediately after completing a short form. The data appears complete.
We would like to thank you for your positive comments about the contribution of the present work.

Specific comments

Section 2.1 Selection of validation sites: This section comes across as a literature review of different methods, and I'm not sure what is actually being implemented in this paper from reading this section. Can you clarify in the paragraph (e.g. does each dataset use a different method?)
This point was also indicated by the anonymous referee # 1 in his/her specific comment (point 1). We have added an explanation in section 2.1 clarifying this aspect. Please note that table 2 of the reviewed manuscript summarizes the sampling method applied in each dataset.

At the end of the Introduction, the overview of the paper is a bit vague. I think this would benefit from a clearer outline of the structure, and a list of the datasets that are considered in this paper to give a better overview up front.
Thank you, we have mentioned at the end of the introduction the datasets included in BARD and clarify the project where they have been produced and the contents of the manuscript.

Line 220 – only data in June to October is considered for this dataset. This covers the main fire season in this region, but how are the fires outside of the fire season dealt with?
BrFLAS dataset has been removed from the database since it does not follow CEOS cal-val standards. Please, see short comment 6 and response.

Presumably the temporal length of the reference files is such that it covers multi-day burning. It is worth pointing this out in the text explicitly.
Reference data include all the fire perimeters occurred between the two dates of the Landsat images used to generate them. This is a standard practice in BA validation. We have added a comment at the end of the section 2.2. and modified figure 5 to clarify this point.

It would be useful to include some text describing how one might use all these different reference datasets in practise. Should they all be used together, and if so how should the range be accounted for?
Thank you for this relevant question. This question is also related to specific comments (point 2) from the anonymous referee # 1, general comments of referee # 3 and SC3 (point4). Datasets are not supposed to be used together, as they have been obtained from different methods, rather users can choose the datasets that best suits their needs. As suggested by referee # 1, we have added the data necessary to make probability estimates of accuracy for those datasets obtained through stratified random sampling (Tables in Appendix A of the reviewed manuscript).

How were these datasets selected? Are there any other datasets available that are not included here, or are these the only ones available? I suggest including some explanation of this in the text.
Yes, there are other datasets that have been produced by other authors (e.g. Boschetti et al. 2016;2019). We made a general announcement through the GOFC-GOLD Fire implementation team list of scientists working on BA products and to our network of fire scientist. The resulting

database includes files those that the authors were willing to share publicly and met the CEOS cal-val standards.

Are all the datasets related to FireCCI? It seems so from the description of the data via the link, but not in the paper.
Only the datasets with the 'FireCCI' word in its name were produced under the FireCCI project, the rest of the datasets come from others projects. We mention this in the introduction: 'These validation files were compiled from different international projects and years…'. In addition, we have added the project name of each dataset in Table 2.

Most modellers use NetCDF, if it would be nice if this format was considered for future releases.
Thank you, we will keep in mind your suggestion for future releases. We don't usually use the NetCDF format for the reference files, but users interested in such format can easily do the conversion from .shp to .nc with the open tool 'ncl_convert2nc' that can be downloaded from '*https://www.ncl.ucar.edu/Document/Tools/ncl_convert2nc.shtml*'.

Technical corrections
References to figures ("Fig.") throughout the text is sometimes with a space and sometimes without
Thank you, we've added a space in those where it was missing.

Line 49 – change to "acquired in the year 2000"
Done.

Line 182 - "consists of"
Done.

##########

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.

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.
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.
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?
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.
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.

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.
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?
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.
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 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?
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.

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.
Thank you, we have modified figure 5 (now figure 3) 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.
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.
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."
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"
Done

L57-58: Acronyms MERIS and MODIS not defined
Done.

L85: change “amount” to “number”
Done.

L91: change “sample” to “samples”
Done.

L213: change “covering the 77%” to “covering 77%”
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”
Done.

##########

David Roy short comment (SC1)
Received and published: 28 July 2020

Table 2 summarizes the number of reference files but does not provide summary information on the total areas of the 3 mapped classes (burned, unburned, no data). Please add a new table providing this information for each project and for all the projects together. This is needed because (i) Landsat and Sentinel-2 images have different areas (_185x185 km and _110x110 km), (ii) different image spatial subsets were mapped (i.e., not the entire image extents) by the different projects, (iii) the "no data" class includes areas where the interpreter did not undertake the mapping and it is unclear if this was a major proportion and/or if it varied among the projects. It would be helpful to then place the total mapped burned and unburned areas in the context of (a) the total global land area and (b) the typical total annual global area burned, and update the abstract and conclusion accordingly.

Thank you for your suggestion. Information about the total area from the 3 mapped classes was already included in the medatada files, however, we didn’t mention it in the manuscript. We have added the proper comment in the .csv metadata file description and included the suggested table summarizing the total area mapped of each dataset and the area of the three mapped categories.

##########

David Roy short comment (SC2)
Received and published: 28 July 2020

Could you explain the long and short units methodology better please. Currently as written I cannot understand it. Figure 5 is helpful but it does not show the case where there are "no data" areas (for example, due to cloud and/or the Landsat SLC-off issue) in the image time series. Please clarify this in the paper text - paying particular attention to how the "no data" pixels are handled in the long unit derivation. I suspect that there are underlying assumptions that reduce the utility of the long unit results for validation. For example, it is well known that in many regions the burn signal dissipates rapidly and that clouds occur commonly and often at the time of Landsat or Sentinel-2 overpass. Thus, to my mind, the long unit may (i) fail to capture the true area burned over the time series, (ii) reduce the proportion of the image area that is mapped as burned and/or unburned. Please clarify and discuss.

We have extended the description of the methodology to create the short and long units in section 2.2 and updated figure 5 (now figure 3) including unobserved areas, we hope it will be clearer now. As we explain in the methodology to build long units, consecutive pairs of images are used in order to avoid burn signal loss within the period covered by the long unit. On the other hand, it is true that may the proportion of the mapped region could be reduced in the spatial dimension, as 'no data' in any of the image pairs is kept into the final reference data. However, this should not affect the suitability of long units as reference data, please note that, for example, in Boschetti et al. (2019) images with cloud cover up to 70% are used for validation. Furthermore, long units have a crucial advantage over short units as they reduce the impact of the temporal reporting accuracy in the accuracy estimates. We consider that both, short and long units, are complementary and useful for validation.

##

L. Boschetti short comment (SC3)
Received and published: 28 July 2020

Given the effort and cost involved in generating validation dataset, the standardization and documentation of existing datasets for future use is certainly a meritorious effort, and there is no doubt that the datasets made available by the authors of this paper will find a use in the fire community.

I have however some concerns.
1) Scope of the paper and qualification of the dataset The way the dataset is presented could lead to some confusion and misinterpretation. The paper title refers to the 'development of a standard database' and the abstract refers to the present work as the compilation of 'the first Burned Area Reference Database'. This is misleading, because the work described in the present paper is limited to the collation of existing datasets, through standard GIS operations described in Section 2.3, namely the conversion of the various datasets to the same file format (shapefile), the use of standardized file names and the creation of simple metadata (Table 1). The wording of the abstract, i.e. referring to BARD as 'the first Burned Area Reference Database', is incorrect, as this is not the first burned area reference dataset - all the datasets that constitute the BARD are pre-existing. Maybe 'the first publicly available burned area reference dataset' would be a more appropriate statement.
We have followed your suggestion and changed the sentence to "the first publicly available burned area reference dataset". Actually, identical sentence was already included in the conclusion section: '*the first publicly available burned area reference dataset'* where we clearly mentioned that '*BARD is the first publicly available database that compiles and standardizes previously generated validation reference data*.'

2) Degree of novelty Section 2.3 is the only section that reports original work (i.e. the conversion of all data to shapefile, the standard filenames and the metadata), while the rest of the methods document what was done by the various research teams in the projects that provided the data.
The novelty of this paper is the compilation, standardization and public release of existing BA validation datasets, never done before.

3) Sampling. Section 2.1 ('Selection of the validation sites') describes a procedure for stratified random sampling of the burned area reference data that was followed by some of the collated datasets (but not all), which is extremely confusing. An inattentive reader might be led to believe that the BARD dataset itself is the result of a stratified random sampling, rather than the collation of datasets some of which were the results of stratified random sampling (albeit with different methods) and some that are not.
Table 2 and the documentation of the database included a clear description to the contents, but we have further clarified this point in section 2.1 to avoid confusion. Please, note that table 2 summarizes the sampling methodology applied to each dataset.

4) Stage 3 validation data set. Much is said throughout the paper of the compliance with CEOS Stage 3 validation requirements, but the BARD dataset as currently defined does not meet those requirements, i.e. it would not allow for the use of unbiased estimators of the accuracy metrics, and their associated standard errors. In the current form, pieces of BARD could be used for a Stage 3 validation, whereas other pieces could only be used for a Stage 1 or Stage 2 validation. Could the authors provide a harmonized statistical framework for the estimation of accuracy metrics from the whole BARD dataset?

BARD is a compilation of datasets that have been produced in different projects where different methods were applied. Even those datasets produced in the throughout the life of the FireCCI project present substantial differences. Please note that FireCCI project is a long term project that started in the year 2010 and, through the years, methods have been improving. That said, the aim of BARD is not to provide a harmonized statistical framework for all contributing datasets, because BARD is not a dataset itself but a compilation of datasets produced by different international projects and years. If that was the aim of BARD, we would only have made available the FireCCI global (2003-2014) dataset which is the one that covers the longest period but, instead, we choose to make all possible datasets available, and leave users the freedom to use the dataset or datasets that best suits their needs.

##########

L. Boschetti short comment (SC4)
Received and published: 28 July 2020

Dataset harmonization It is apparent that the individual datasets collated in BARD were derived using a variety of semi-automatic procedures, and in the context of projects that had a somewhat different emphasis. How were the data harmonized so that they can be used together meaningfully? The title, abstract, introduction and conclusion imply a degree of harmonization between the datasets that is well beyond what was done, and might be seen as overstating the potential of the BARD to be a 'standard database'. Furthermore, there is no formal discussion in the paper of what requirements/ criteria/standards should be met by a 'standard database'.

We have answered this comment in point 4 of SC3. The standardization refers to the formats and documentation of the contributing datasets. This is something similar to what has been done in other databases made publicly available, as it is the case of a recent paper by Yebra et al. (2019) with fuel moisture content measurements.

##########

Vitor Martins short comment (SC5)
Received and published: 4 August 2020

The standard datasets are relevant for validation of burned areas and the scientific community will be interested in such product. However, I found quality issues that limit the application without additional evaluation of files. Since these datasets have certain degree of automation in the production, further inspection is required to guarantee the high-quality in reference files. For instance, some areas present no-data/unobserved labels without a clear reason. When the reference dataset omits complex burned areas, the validation results tend to be biased. Other problems were observed in water pixels labeled as "unburned", and harvest areas as "burned". The authors should acknowledge the problems in these reference files and improve the quality as much as possible.

As burned perimeters of reference files were retrieved using semi-automatic classifications technics, all the reference files were visually inspected by an experienced interpreter and double checked (or triple checked in the case of the BAECV CONUS

(1988-2013) dataset) by another independent interpreter and the errors detected in the initial classification were manually edited until no errors were found. We mentioned this procedure throughout the manuscript in lines 76, 107, 177, 201-204, 218, 258 and 294. There are two main situations where you can find 'no-data/unobserved labels without a clear reason': the first one, is when the pre- or post-fire image (or both of them) used to retrieve the burned area has a region covered by clouds and shadow clouds, and the cloud/shadow mask applied to the images does not properly mask them. This situation makes difficult the correct classification of the pixels located near the clouds and could lead to an incorrect classification. To avoid this issue, cloudy regions of the image are excluded by using a mask manually created when necessary. This does not imply a reduction of the quality of the reference perimeters but a reduction of the interpreted area. The second one, is related mainly with crop areas, where harvested areas could be classified as burned areas as you point out in your comment. In some regions, especially those with dark soils, is very hard to differentiate between harvested crops and burned areas. Despite this issue, we made a great effort to interpret those areas and used some strategies to minimize the errors in those cases. In this sense, active fires from MODIS and observable flames on the images can help to identify which crops have burned and use only those pixels to train the classifier. However, there are situations where the classification results are quite uncertain and masking those areas as unobserved is preferable. Respect the water pixels labelled as 'unburned' issue, it has to do more with the established criteria in the validation methodoly and not with labelling errors. Masking water as unobserved or no data could hide commission errors of coarse resolution BA products, especially in those regions where a large number of small water bodies surface cover a significant part of the validation area (e.g. Boreal and Tundra biomes). Labelling water as unburned was the criterion adopted in the FireCCI datasets and others reference datasets of BARD, we are aware that may this criterion is not the optimal for all the regions of the world but this is a question that requires further research to know exactly the impact of such decisions.

We acknowledge that reference files will have always a certain degree of uncertainty due to the remote sensing limitations but we consider that the reference files compiled in BARD are the best approximation to the ground truth that allows the current technology.

##########

Vitor Martins short comment (SC6)
Received and published: 5 August 2020

I examined all the Brazilian (BrFLAS) data including comparing them to the multi-date Landsat images they were derived from. Two obvious issues:
1) None of the Brazilian data have a "no data\unobserved" class. This would only be correct if the images were always cloud- and shadow- free and but this is not the case.
For example, see below.
2) There are burned areas that are not mapped as "burned" because one of the images
was cloud/shadow obscured. However, incorrectly, they have not been mapped as "no data\unobserved" (for example, see in red circle below). This makes these data difficult to use for validation, or as a reliable source of training data for classification purposes (as without looking at the images I would assume incorrectly that these areas were unburned).

Thank you very much for show us this issues about the BrFLAS dataset, we really appreciate the time dedicated to check the accuracy of the data. We agree with your

observations about the BrFLAS dataset and we have decided to exclude this dataset from BARD until these issues are fixed.

Development of a standard database of reference sites for validating global burned area products

Magí Franquesa[1], Melanie K. Vanderhoof[2], Dimitris Stavrakoudis[3], Ioannis Z. Gitas[3], Ekhi Roteta[4], Marc Padilla[5], Emilio Chuvieco[1]

[1]Environmental Remote Sensing Research Group, Department of Geology, Geography and the Environment, Universidad de Alcalá, Calle Colegios 2, Alcalá de Henares, 28801, Spain
[2]U.S. Geological Survey, Geosciences and Environmental Change Science Center, P.O. Box 25046, DFC, MS980, Denver, CO 80225, United States
[3]Laboratory of Forest Management and Remote Sensing, School of Forestry and Natural Environment, Aristotle University of Thessaloniki, P.O. Box 248, GR-54124, Greece
[4]Department of Mining and Metallurgical Engineering and Materials Science, School of Engineering of Vitoria-Gasteiz, University of the Basque Country UPV/EHU, Nieves Cano 12, Vitoria-Gasteiz, 01006, Spain
[5]Centre for Landscape& Climate Research, Department of Geography, University of Leicester, Leicester LEI 17RH, United Kingdom

Correspondence to: Magí Franquesa (magin.franquesa@uah.es)

Abstract. Over the past two decades, several global burned area products have been produced and released to the public. However, the accuracy assessment of such products largely depends on the availability of reliable reference data that currently do not exist on a global scale or whose production require a high level of dedication of project resources. The important lack of reference data for the validation of burned area products is addressed in this paper. We provide the first publicly available Burned Area Reference Database (BARD) that was created by compiling existing reference BA datasets from different international projects. BARD contains a total of 2,661 reference files derived from Landsat and Sentinel-2 imagery. All those files have been checked for internal quality and are freely provided by the authors. To ensure database consistency, all files were transformed to a common format and were properly documented by following metadata standards. The goal of generating this database was to facilitate BA algorithm developers and product testers reference information that would help to develop or validate new BA products. BARD is freely available at: https://doi.org/10.21950/BBQQU7 (Franquesa et al., 2020).

1 Introduction

Validation is defined by the Committee on Earth Observing Satellites Working Group on Calibration and Validation (CEOS-WGCV) as "the process of assessing, by independent means, the quality of the data products derived from the system outputs" (CEOS-WGCV, 2012). Validation helps in evaluating the utility and limitations of using any remote sensing (RS) product, particularly on whether user accuracy requirements are met. For this reason, validation should be part of any RS project, even though it requires additional effort and cost that is not aimed at improving accuracy but rather to measure it. Validation implies comparing our results to reference data, assumed to represent the actual conditions of the target variable at the satellite overpass

Eliminado: Renata Libonati[3,4], Julia A. Rodrigues[3], Alberto W. Setzer[5],

Eliminado: es

Eliminado: s

Eliminado: burned area

Eliminado: The Database

Eliminado: 2769

Eliminado: burned area

Eliminado: or

Eliminado: reference

Eliminado: This should help future users of this database to read and convert the files to their own preferred formats or projections. The database

time. In the case of global studies, it is very difficult to generate reference data for the wide variety of planetary conditions, thereby complicating validation. Some of the global variables (e.g. temperature and surface radiation) can be validated from ground sensor networks, such as weather stations, buoys or Aerosol Robotic NETwork (AERONET) sensors. Other variables are more difficult to validate, as they require generating global reference data that are based on higher-resolution sensors than those used to obtain the global product. This is the case of land cover or burned area products, which require first designing a sample strategy using statistically valid protocols and then extracting from the selected sites the reference polygons to be compared with the global datasets. Despite the time and effort required to derive reference datasets, accuracy assessment is a critical part of any global RS project and making these reference datasets publicly available will facilitate product comparison and lower the burden of validating future products.

Several global burned area (BA) products have been produced in the last two decades, providing an estimation of fire activity worldwide (Chuvieco et al., 2019). The first of these products was the Global Burned Area (GBA2000), based on daily VEGETATION (VGT, 1 km resolution) images acquired in the year 2000 and was generated by the Joint Research Centre of the European Union (Grégoire et al., 2003). The same year, the European Space Agency developed the GLOBSCAR BA product, also at 1 km^2, derived from daytime ERS-2 (European Remote Sensing Satellite) ATSR-2 (Along Track Scanning Radiometer) data (Simon et al., 2004). Other 1 km resolution global BA products released by European projects include the L3JRC (Tansey et al., 2008) covering the period from 2000 to 2007; GlobCarbon (Plummer et al., 2006), produced from 1998 to 2007; and the Copernicus GIO_GL1_BA products. These three products were derived from VGT images, although in the GlobCarbon project, ATSR images were used as well. More recently, the FireCCI (Climate Change Initiative) project (https://esa-fire-cci.org, last access: 25 March 2020), part of the European Space Agency (ESA) CCI programme, has generated three global BA products, based on Medium Resolution Imaging Spectrometer (MERIS) at 300m resolution (FireCCI41: Alonso-Canas and Chuvieco, 2015) and Moderate Resolution Imaging Spectroradiometer (MODIS) 250m data (FireCCI50: Chuvieco et al. (2018) and FireCCI51: Lizundia-Loiola et al. (2020)). NASA (National Aeronautics Space Administration) released in mid-2008 the MCD45A1 product derived from 500 m MODIS imagery (Roy et al., 2008), which has now been superseded by MCD64A1 at the same resolution but with a different BA algorithm approach (Giglio et al., 2009; 2018).

These global BA products have been validated by comparing them with reference data generated from medium resolution sensors (such as those on board the Landsat, SPOT (Satellite Pour l'Observation de la Terre), or Sentinel-2 missions). These reference data were typically derived from multitemporal pairs of images to properly date the validation period.

According to the representativeness of samples used to perform product validation, the CEOS-WGCV Land Product Validation (LPV) subgroup defined four validation stages with the level of sampling effort and statistical rigor increasing at each stage (https://lpvs.gsfc.nasa.gov/, last access: 25 March 2020). 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 (Tansey et al., 2004; Roy et al., 2005). Roy and Boschetti (2009), for instance, reported validation results for the MCD45A1 product using a set of 11 Landsat scenes distributed across southern Africa. Chuvieco et al. (2008)

Eliminado: sensors

Eliminado: is

Eliminado: product efforts

Eliminado: global BA product

Eliminado: ,

Eliminado: was

Eliminado: for

Eliminado: at coarse resolution

Con formato: Superíndice

Eliminado:

Eliminado: Both BA products, GBA2000 and GLOBSCAR BA, had a nominal pixel size of 1 km^2.

Eliminado: All t

Eliminado: (

Eliminado:)

Eliminado: 's

Eliminado: Climate Change Initiative

Eliminado: MODIS

Eliminado: , the latter at 250m spatial resolution.

Eliminado: 2010

Eliminado: files

Eliminado: without probabilistic meaning (i.e. direct sampling),

Eliminado: and

Eliminado: main

validated a regional product for Latin America using 19 Landsat scenes and 9 China–Brazil Earth Resources Satellite (CBERS) scenes that were donated by regional space agencies when access to the Landsat archive was not yet free and open to the public, thereby limiting the number of selected validation sites. The MCD64A1 Collection 5 was not formally validated, and the most recent MCD64A1 Collection 6 products were first validated using a set of 108 Landsat scenes distributed across a wide range of fire-affected ecosystems but not selected via probability sampling (Giglio et al., 2018). 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). Previous statistical validation of NASA and FireCCI BA products were conducted by Padilla et al. (2014; 2015) using a set of 105 randomly selected Landsat scenes for a single year (2008) and by Chuvieco et al.(2018) using a multitemporal reference dataset of 12 years. Other projects covering large areas have been developed in the USA using Landsat data across six years (Vanderhoof et al., 2017) and Africa using Sentinel-2 Multispectral Instrument (MSI) images (Roteta et al., 2019), where validation sites were selected through probability sampling. In all cases, reference datasets were created based on independent interpretation of BA, controlled by visual inspection. The importance of applying probability sampling to collect reference data has been highlighted by different authors as a critical feature of the sampling design protocol to achieve statistically rigorous assessment (Stehman, 2001; 2009; Olofsson et al., 2014; Stehman and Foody, 2019). Thus, in contrast to such reference data collected by convenience, ease of access, or other methods that lack randomization, data collected through probability sampling makes it possible to obtain rigorous estimates of accuracy.

The main bottleneck for validating global BA products or global BA algorithms is the generation of reference BA datasets. To facilitate the activity of BA algorithm developers, this paper aims to present and deliver to the scientific community the Burned Area Reference Database (BARD), a set of reference BA perimeters that can be used as reference data for validation of BA products or to help the development of BA algorithms (obviously, the same files cannot be used for both training and validating an algorithm). These validation files were compiled from different international projects and years, therefore the resulting database will facilitate the assessment of BA algorithms in a wide range of ground conditions.

The BARD includes the following datasets of reference data: FireCCI global (2008), FireCCI global (2003-2014), FireCCI Africa (2016), FireCCI Africa S2 (2016) that were produced within the framework of the FireCCI project; the CONUS (contiguous United States) Landsat Burned Area (1988-2013), developed within the Landsat Level-3 Science Products project, and NOFFi Greece (National Observatory of Forest Fires, 2016-2018) that were produced within the NOFFi project.

The paper presents the methods that were used to generate the BA reference data paying particular attention to the sampling design and reference data retrieval methods applied to the different BARD datasets. The data specifications to transform all the files to a common standard format and file structure are then presented. Finally, a detailed description of each dataset included in BARD is provided and the main dataset features are then summarized to facilitate a general overview.

Eliminado: full spatio-temporal

Eliminado: ,

Eliminado: have been

Eliminado: 130 images

Eliminado: . In both cases, reference datasets were created based on independent interpretation of burn perimeters, controlled by visual inspection

Eliminado: facilitating

Eliminado: test

Eliminado: The paper presents the methods that were used to generate the BA reference perimeters and then to transform all the files to a common standard format and file structure.

2 Methods

2.1 Selection of validation sites: sampling design

High-quality reference data generation is an expensive and time-consuming task, which constrains the total number of validation sites that can be established in any validation exercise. For this reason, sampling design is critical to make the most of the resources available and ensure the highest precision of accuracy estimates given the available resources to generate reference data. Padilla et al. (2014; 2015) implemented a stratified random sampling design that allowed for global BA accuracy inferences for the first time. Boschetti et al. (2016) extended the sampling design to include the temporal dimension of the sampling units. More recently, Padilla et al. (2017) presented a first approach to efficiently stratify the population and allocate the samples across strata. Chuvieco et al. (2018) conducted a multi-annual accuracy assessment across 12 calendar years (2003-2014), reporting for the first time the temporal accuracy variation of global BA products. Meanwhile, Boschetti et al. (2019) validated the MCD64 c6 BA product, but instead of using the calendar year, the authors used a fire year (from March 1st 2014 to March 19th 2015) as defined in Boschetti and Roy (2008).

The sampling design protocols to validate BA products were therefore developed considering the rarity and ephemeral nature of the BA, which is indeed a special case of land-cover change (Stehman and Foody, 2019). When selecting samples for obtaining probability inferences, the allocation of samples should follow a probability sampling design, to compute unbiased population estimates. For BA product validation, this implies selecting samples considering the spatial and temporal dimension. The spatial dimension of sampling units is usually defined by the Thiessen scene areas (TSAs) constructed by Cohen et al. (2010) and Kennedy et al. (2010) specifically for use with Landsat WRS-2 frames (Worldwide Reference System, Fig. 1a). The key advantage of TSAs is that they provide non-overlapping Landsat-like frames, which allow for a convenient computation of unbiased estimators (Gallego, 2005). The temporal dimension of sample units is defined by the acquisition dates of the pre- and post-fire images. For example, in Boschetti et al. (2019), the validation period (1 year) was divided into equal temporal size sampling units using the 16-day Landsat 8 acquisition interval, thus allowing for the temporal random selection of the reference images. This temporal partitioning, also makes it possible to intensify the sample in strata that comprise the fire season and where burning is more likely to occur (Stehman and Foody, 2019). However, longer period intervals (>100 days) are used to define sampling units to allow a long temporal overlap of reference data with the BA product, which helps to disentangle the spatial errors from the temporal errors of the BA product (Roteta et al., 2019; Lizundia-Loiola et al., 2020).

In any case, sample units are then stratified to properly represent the variety of conditions that affect the accuracy of BA products. This stratification is usually based on (a) major Olson biomes (Olson et al., 2001) (Fig. 1b) and (b) the BA extent provided by a global BA product considered to be reliable or active fire detections, assigning each sample unit to high or low BA strata based on a threshold that can be specifically adapted to each biome stratum as in Padilla et al. (2017) or simply set as the 20th quantile of the cumulative distribution of active fire counts as in Boschetti et al. (2016; 2019).

Eliminado: amount

Eliminado: improved

Eliminado: by specifically including

Eliminado: at

Eliminado: sample

Eliminado: A

Eliminado: for example a stratified random sampling, allows for

Eliminado: and is commonly defined by

Eliminado: s

Eliminado: are

Eliminado: consecutive

Eliminado: global BA product

Eliminado: a

One of the advantages of the stratified sampling design adopted for BA maps validation previously mentioned was that it allows for rigorous estimates of global BA accuracy. However, 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 the 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.

Stratified random sampling design was applied to several datasets included in BARD: FireCCI global (2008), FireCCI global (2003-2014), FireCCI Africa (2016) and the CONUS Landsat Burned Area (1988-2013). FireCCI Africa S2 (2016) was obtained also by probability sampling but, in this case, applying a systematic sampling design. NOFFi Greece (2016-2018) is the only dataset of BARD that was obtained through convenience sampling rather than probability sampling.

To report BA accuracy from these stratified sample datasets, users should apply the proper estimation formulas detailed in the associated articles (see Table 2) and use the additional information as the stratum of each sampled unit and the stratum sizes of the stratified sampling, provided in the metadata files and tables of appendix A, respectively.

2.2 Reference data generation methods

Following the recommendations of the CEOS Calibration/Validation group, all the burned perimeters of BARD were derived from multitemporal comparison of medium resolution satellite imagery (Landsat TM (Thematic Mapper)/ETM+ (Enhanced Thematic Mapper plus)/OLI or Sentinel-2 MSI). Burned patches included in the files are only those that occurred in between the two satellite images used to generate the reference data (Fig. 2). The procedures implemented to obtain those burned patches are diverse, depending on the dataset, but all include a semi-automatic procedure (e.g. Bastarrika et al., 2011) and then a visual inspection to confirm that the detected perimeters were actually burned areas. In some cases, the semiautomatic classification was enhanced with polygons manually digitized. In several cases, this visual inspection was confirmed by another interpreter to double check the quality. When parts of the scene could not be observed or interpreted because of clouds or sensor problems (i.e. Scan Line Corrector (SLC)-off problems of ETM+), either in the pre- or post-fire images, they were classified as no-data. This was done to make sure that only areas with reliable data were included in the reference files. Regarding 'unburned' category of reference data, different criteria were applied to label seas and inland water bodies in the different datasets. Thus, for FireCCI global (2008), FireCCI global (2003-2014), FireCCI Africa (2016) and CONUS Landsat Burned Area (1988-2013) datasets, surface waters were classified as 'unburned' while in FireCCI Africa S2 (2016) and NOFFi Greece (2016-2018), the 'no-data' category was applied to label them.

It should be noted that reference data are not just high accuracy BA products generated by well-designed algorithms using medium- or high-resolution imagery. Rather, reference data following international standards should provide reliable burned area but also the unburned surface of the interpreted geographic region and the unobserved/unmapped areas within the region, as shown in Fig. 2c.

Eliminado: reference
Eliminado: -
Eliminado: reference files
Eliminado: files
Eliminado: procedure
Eliminado: is
Eliminado: Optionally
Eliminado: can be overwritten by
Eliminado: manually
Eliminado: Parts
Eliminado: that
Eliminado: cannot
Eliminado: , either by
Eliminado: by
Eliminado: validation
Eliminado: process

Like the sampling units from which reference data are derived, reference data can be defined by its spatial and temporal dimension. The spatial dimension is a function of the geographic extent interpreted to obtain the reference data, where the size varies depending on the criteria adopted in each project. For example, reference data from the FireCCI global (2003-2014) dataset were spatially defined by a frame of 30 x 20 km located at the centre of the Landsat images, whereas the entire Landsat scenes were used in the case of the CONUS Landsat Burned Area (1988-2013) dataset. The spatial extent used in the datasets included in BARD will be specified in section 2.4 where a detailed description of each dataset is provided.

The temporal dimension of the reference data represents the period defined by the acquisition date of the pre- and post-fire images used to generate them. Regarding the temporal length of the reference data, the FireCCI project adopted the terms 'short unit' (SU) and 'long unit' (LU). The former refers to those reference data derived from a pair of consecutive images separated by 16 days or less (the temporal span between two Landsat acquisitions). The latter is defined by a series of consecutive SUs covering at least 100 days. LUs allow for long temporal overlaps between validation and product data, reducing or minimizing the impact of the product's temporal reporting accuracy in the accuracy estimates (Padilla et al., 2018). The combined use of SUs and LUs is useful to assess such and contextualize impact. A LU BA map consists in the combination of consecutive SU maps (Fig. 3). A pixel classified as no-data in any of the SU maps is kept as such in the LU BA map. This is to ensure that any pixel available data is observed frequently (every 16 days or less) and an eventual burn is not missed due to simply a fast recovery of the vegetation. The permanently observed pixels, were classified as burned in the LU if they were detected as burned in any SU of the time series covered by the LU. The presence of no-data (e.g. due to clouds) in a single image may reduce drastically the spatial cover of available data in the resulting LU. Therefore, BA maps are generated for every single SU, but the BA map for a LU is generated by accumulating the consecutive SUs of the same TSA. The length of the LU would depend on the existing cloud-free consecutive SUs. For example, if 8 consecutive SUs, all covering the same temporal length (e.g. 16 days) are cloud free and the 9th image has 90% of the area cloud covered, the LU would include only the first 8 SU maps, even if SU were generated for the 9th and 10th consecutive images.

As burning is detected on any given single image in between the period covered by two satellite acquisitions, all burned patches are dated based on the second reference image of a multitemporal pair. Therefore, SUs will have the same date for all the burned patches, while LU reference data will have burned patches from different dates as multiple pairs of images are used to build the LU (Fig. 3).

Among the datasets included in BARD, SUs were used in the FireCCI global (2003-2014) dataset as part of the sampling design, and LUs were used for the FireCCI Africa (2016) dataset. Reference data from the rest of the FireCCI project datasets (FireCCI global (2008) and FireCCI Africa S2 (2016)) and CONUS Landsat Burned Area (1988-2013) dataset, were retrieved from a single pair of images with a variable time lapse between pre- and post-fire images. Thus, the temporal length of those reference data was determined by the availability of suitable images and the duration of the burned signal. The NOFFi Greece (2016-2018) reference data were obtained considering a time-series of Sentinel-2 images, but with variable length and non-consecutive time-series step.

2.3 Data specifications

Each dataset of BARD is organised in three folders with associated files including: (a) 'metadata', which contains a .csv file containing the file name of all the reference files included in the dataset, along with additional information such as the temporal length (days), the total number of images interpreted (n_images), the area (m^2) of each mapped category ('burned', 'unburned' and 'unobserved'), the land surface and total area of each reference data file. For those datasets where a stratified random sampling design was used, the .csv file also specifies the stratum of each sampled unit and the size (tsa_area) of the corresponding TSA; (b) 'regions', which contains an ESRI shapefile (*.shp) containing all the sample sites (TSAs or Sentinel-2 tiles) covered by the dataset; and (c) 'shapefiles', containing the validation reference shapefiles ordered by year. They are also released in shape (.shp) format.

All datasets are in UTM/WGS84 projection. The name of the files is defined as follows: 'Project_RD_ppprrr_yyyymmdd_yyyymmdd' (e.g. FireCCI_RD_164069_20160514_20160709'), where:

Project = Project in which the reference data were generated.

RD = stands for Reference Data.

ppprrr = refers to the Landsat Worldwide Reference System (WRS) path (ppp) and row (rrr) of the scene. For collections where Sentinel-2 was used instead of Landsat images, ppprrr refers to the Sentinel-2 tile (e.g. FireCCI_RD_T28PET_ 2016011 1_20160311').

yyyymmdd (year, month, day). The first date corresponds to the pre-fire date, which is the date of the first image used for BA detection; the second one refers to the post-fire date, which is the date of the last image used for generating the reference fire perimeters.

The following attribute fields are included in the shapefiles (Table 1):

- category:
  - 1: Burned area. This category includes all polygons detected as burned
  - 2: No-Data. This category includes all polygons that could not be interpreted or were not observed by the sensor, either by clouds and/or cloud shadows, topographic shadows, smoke, or sensor errors (for instance, those caused by SLC-off problems of ETM+ after May 31, 2003).
  - 3: Unburned. This category includes all polygons observed as not burned within the limits of the area covered by the image.
- preDate: Acquisition date of the image taken before the occurrence of the fire: yyyy-mm-dd (year, month, day).
- postDate: Acquisition date of the image taken after the fire: yyyy-mm-dd (year, month, day).
- preImg and postImg: The pre- and post-fire Landsat scene identifier (e.g. 'LC80260422013124LGN01'). For reference files based on S2 images, the datastrip ID is used instead.
  (e.g. 'S2A_OPER_MSI_L1C_TL_SGS__20160420T171415_A004324_T28PEB_N02.01').

Eliminado: Existing validation datasets from different global and regional projects were compiled in the Burned Area Reference Database (BARD).

Eliminado: such

Eliminado: like

Eliminado:) or

Eliminado: for each reference

Eliminado: Reference files are released as Esri shapefile (*.shp) format

Eliminado: in

- path: The Worldwide Reference System-2 (WRS-2) path of the Landsat scene. For reference files based on S2, the tile number was used.
- row: The row of the Landsat scene. For reference files based on S2, the tile number was used.
- year: The year of the validation dataset.
- area: Area in square meters (m^2) calculated on the WGS84/UTM Cartesian plane.

2.4 Reference datasets

2.4.1 FireCCI global (2008)

The FireCCI global 2008 reference dataset was created using a stratified random sampling design ((Padilla et al., 2014; 2015), Table A1). Two levels of spatial stratification were used to select the spatial units based on TSAs derived from the Landsat World Reference System 2 (WRS-2). Spatial units were first stratified across seven aggregated Olson biomes (Olson et al., 2001). Each biome was stratified into high and low BA extent based on the Global Fire Emissions Database (GFED) Version 3 (Giglio et al., 2009; 2010). A total of 101 images from Landsat-5 TM and 109 for Landsat-7 ETM+ satellite sensors were used to retrieve BA perimeters. The complete scene was used for Landsat-5 TM images, whereas only the centre of Landsat-7 ETM+ scenes were interpreted in order to avoid data SLC gaps. BA perimeters were derived using a semi-automatic algorithm developed by Bastarrika et al. (2011), where high burn severity pixels were selected to train core burned area, and adjacent lower burn severity pixels were added to the core detected patches using a region-growing algorithm.

The FireCCI global 2008 dataset includes 105 reference data files, derived from single pair of images, for the year 2008. The temporal length of reference data varies between 8 and 144 days: 79% of image pairs were separated by 32 days or less, 16% between 32 and 100 days, and 5% by more than 100 days with a maximum time gap between the pre- and post-fire image of 144 days. The total area of reference data is $1.76 \cdot 10^6$ km^2, of which 1.35% corresponds to burned category, 88.35% to unburned and 10.30% to unobserved category. The location and temporal length of the reference data is shown in Fig. 4. This reference dataset is compliant with CEOS-LPVS Stage 3.

2.4.2 FireCCI global (2003-2014)

The FireCCI global (2003-2014) dataset covers a period of 12 years, from 2003 to 2014 (Padilla et al., 2018), and was generated in the framework of the FireCCI project with the collaboration of the Copernicus Global Land Service (CGLS). The reference data were derived from consecutive Landsat images separated by 8-16 days for each selected TSA and year. A total of 585 images from Landsat-5 TM, 1564 from Landsat-7 ETM+ and 209 from Landsat-8 OLI (n= 209) satellite sensors were used to retrieve BA perimeters. The sampling units were selected following a stratified random sampling design (Table A2). The total population of sample units were defined spatially by TSAs and temporally by the dates of Landsat images available, filtering out those with a cloud cover greater than 30%. For each calendar year, the sample units were stratified by Olson biomes (Olson et al., 2001) and BA based on MCD64A1 (Giglio et al., 2009). The threshold used to assign the high/low BA strata was defined

Eliminado: The FireCCI global 2008 reference dataset (Padilla et al., 2014) was created using a stratified random sampling design.
Eliminado: II
Eliminado: II
Eliminado: (n=127)
Eliminado: (n=131) images
Eliminado: er
Eliminado: selected to apply
Eliminado: to the previously identified burned areas
Eliminado: 129
Eliminado:
Eliminado: pairs
Eliminado: files
Eliminado: 64
Eliminado: 28
Eliminado: 8
Eliminado: files
Eliminado: 3
Eliminado: This
Eliminado: twelve
Eliminado: . The reference files were generated
Eliminado: 2358
Eliminado: (n=585)
Eliminado: (n=1564)
Eliminado: was
Eliminado: to each sample unit

separately for each year and biome. Once the strata were defined by year-biome-BA, a set of 100 sampling units were selected for each calendar year applying a sample allocation according to Eq. (1):

$$n_h \propto N_h \, \overline{BA}_h \quad (1)$$

where n_h is the sample size to be selected in stratum h, N_h is the stratum size and $\overline{BA}_h$ the BA mean in stratum h.

Finally, a spatial subset window of 30 x 20 km located at the centre of the images was applied for interpretation and BA reference data retrieval. The reference perimeters were extracted from a dedicated Random Forest algorithm, trained for each sampling site, and output maps were visually inspected by two interpreters (Padilla et al., 2018).

The FireCCI global (2003-2014) dataset includes 1200 reference data files from 722 different TSAs and 12 years, from 2003 to 2014. The temporal length of reference data varies between 8 and 16 days. The total area of reference data is $0.72 \cdot 10^6$ km^2, of which 3.85% corresponds to burned category, 71.85% to unburned, and 24.29% to unobserved category. The location and total number of reference data in each TSA are shown in Fig. 5. This reference dataset is compliant with CEOS-LPVS Stage 3.

2.4.3 FireCCI Africa (2016)

The FireCCI Africa reference dataset consists of LU BA maps and was generated for the year 2016 from Landsat imagery (Padilla et al., 2018). It was also generated in the framework of the FireCCI project with the collaboration of the CGLS. The sampling was designed with long units and it was similar to that for the FireCCI global (2003-2014) dataset, as mentioned in the previous section (Table A3). The only difference was the sample size, 50 units instead of 100 units per year. Note that each unit here is much larger, as it consists of multiple image pairs. Two reference perimeter datasets are released: (a) Reference data at SU level, 1052 files with 8-16 day BA maps; and (b) Reference data at LU level, 50 files. The temporal length covered at each LU varies from 24 to 256 days (Fig. 6b): 18% of the LUs cover a temporal length below 50 days, 34% between 50 and 100 days, and 48% are above 100 days. As mentioned in Section 2.2., LUs were defined to be at least 100 days long, although the presence of clouds reduced the actual temporal periods with available data. The total area of LU reference data is $0.023 \cdot 10^6$ km^2, of which 15.72% corresponds to burned category, 49.61% to unburned, and 34.67% to unobserved category. The location, number of image pairs, and temporal length of the LUs reference data are shown in Fig. 6. This reference dataset is compliant with CEOS-LPVS Stage 3.

2.4.4 FireCCI Africa S2 (2016)

The FireCCI Africa S2 BA reference dataset was created to perform an initial validation assessment of the Small Fire Database Fire_cci v1.1 product (FireCCISFD11) produced for the year 2016 for the whole Sub-Saharan Africa (Roteta et al., 2019). Reference data were generated from the comparison of two Sentinel-2 MSI images at 20 m resolution per reference site. Systematic sampling was used to select 52 validation sites based on Sentinel-2 tiles (110 x 110 km) over Sub-Saharan Africa. BA was mapped with the BAMS methodology, which is a semi-automated algorithm (Bastarrika et al., 2014). In short, training polygons for the burned category were defined in each tile, and burned seeds were detected. Then, burned pixels were grown

Eliminado: proportional
Eliminado: = N_h
Eliminado: W
Eliminado: center
Eliminado: twelve
Eliminado: files
Eliminado: files
Eliminado: is
Eliminado: 4
Eliminado: in consecutive 8-16 days
Eliminado: . The sampling was designed with sampling units long in its time dimension, hereinafter referred to as long units as opposed to the 8-16 days' short units mentioned in the previous section. Reference data over long units allows for long temporal overlaps between validation and product data, and among other analysis it allows to assess the effect of product temporal errors on the accuracy estimates. Reference maps at long units consist simply in the concatenation of consecutive 8-16 days' maps (Fig. 5). The sampling design
Eliminado: on
Eliminado: on
Eliminado: perimeters
Eliminado: short units
Eliminado: days
Eliminado: long units
Eliminado: by
Eliminado: long unit
Eliminado: long units
Eliminado: short units
Eliminado: long units'
Eliminado: files
Eliminado: files
Eliminado: A systematic
Eliminado: 100
Eliminado: 100
Eliminado: Burned areas
Eliminado: were
Eliminado: regions

out from these seeds until all pixels for each burned patches were detected. The results were visually analysed to determine the accuracy of the classification and new training polygons were defined if needed. This was done sequentially until all burned areas were mapped and no commission or omission errors were visually detected. Finally, if there was noise created by unmasked clouds and cloud shadows, it was edited and removed manually.

The temporal length of the reference data varies between 10 and 120 days: 86% of the pairs of images were separated by less than 50 days and 14% by more than 50 days with a maximum time lapse of 120 days. The total area of reference data is $0.63 \cdot 10^6$ km^2, of which 8.87% corresponds to burned category, 72.42% to unburned, and 18.71% to unobserved category. The location and temporal length of the reference data are shown in Fig. 7. This reference dataset is compliant with CEOS-LPVS Stage 1.

2.4.5 CONUS Landsat Burned Area (1988-2013)

CONUS Landsat Burned Area (1988-2013) reference dataset (Vanderhoof et al., 2017; 2020) extends across the contiguous United States (CONUS) and was generated to validate the Landsat Burned Area product (Hawbaker et al., 2017; 2020). The sampling design was adapted from the methods used by the ESACCI FireCCI project. Existing FireCCI validation TSAs (n=9) within CONUS were augmented with an additional 19 TSAs for a total of 28 TSAs. The TSAs were stratified across the major Olson biomes (Olson et al., 2001) including (1) temperate forest, (2) Mediterranean forest, (3) temperate grassland and savannah, (4) tropical and subtropical grasslands and savannah, and (5) xeric/desert shrub. TSAs selected within each biome were meant to represent high and low burned areas as specified by the Global Fire Emissions Database (GFED) version 3 (Table A4). Systematic sampling was applied to select 6 validation years spaced out in 5-year increments (2013, 2008, 2003, 1998, 1993 and 1988).

A total of 269 images from Landsat-5 TM, 10 from Landsat-7 ETM+, and 56 from Landsat-8 OLI were used to derive the BA extent. Landsat reference images were limited to those with a geometric Root Mean Square Error (RMSE) < 10 m, <20% cloud cover, and available as a L1T Surface Reflectance product. Time lapse between images was not limited to 16 days and only two images (pre- and post-fire) were used to retrieve BA reference data for each validation site and year. 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. Peak fire season was determined using the distribution of total burned area by month as derived from the MCD45 burned area product (2001-2015). The FMask from the Landsat surface reflectance product was applied to mask clouds, cloud shadows, snow and open water from each image used (Zhu and Woodcock, 2014). For Landsat-7 ETM+ images, SLC off pixels were masked. The low-, medium- and high-intensity development classes (i.e. urban areas) were masked 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.

Eliminado: the whole

Eliminado: covered

Eliminado: consecutively

Eliminado: files

Eliminado: files

Eliminado: BrFLAS Brazil (2015

Eliminado: The BrFLAS Brazil BA reference files were generated by a joined initiative by the Laboratory for Environmental Satellite Applications (LASA) from the Federal University of Rio de Janeiro (UFRJ) and the National Institute of Space Research (INPE) under the scope of the Brazilian Fire-Land-Atmosphere System (BrFLAS) Project (http://idlcc.fc.ul.pt/brflas/index.html, last access: 25 March 2020). The BrFLAS Brazil (2015) dataset includes 84 reference data files for the year 2015 (Rodrigues et al., 2019) covering the 77% of the Cerrado Brazilian biome's surface. The dataset was derived from images mapped every 16 days at a spatial resolution of 30 m using multispectral images from the OLI sensor aboard the Landsat-8 satellite. The BA perimeters were systematically generated using a classification method based on the Normalized Difference Vegetation Index (NDVI) and Normalized Burn Ratio Long-shortwave infrared variation (NBRL) indices, and on the difference of those indices between the post- and pre-images (Melchiori et al., 2014). These burned scars automatically generated were then subject to an independent analysis and visual photo interpretation, including a series of quality control procedures for removing data of reduced accuracy to ensure consistency among all the burn scar samples. Images with cloud coverage greater than 10% were discarded for the analysis and only images from June to October were interpreted to reduce cloud contamination and rainfall episodes. A detailed visual analysis of cloud distribution in pre- and post-fire images was performed to avoid commission errors. Finally, all the mapped burned scars retrieved from the 16 days' consecutive pair of images were merged for each path/row Landsat-8 scene and the corresponding unburned region was added to build the final refere ...

Eliminado: . A systematic

Eliminado: 336

Eliminado: (n=269)

Eliminado: (n=10)

Eliminado: (n= 56)

Eliminado: image

Eliminado: perimeters

Eliminado: path/row

Eliminado: out

Eliminado: out

Eliminado: out

Eliminado:)

Eliminado: (

Eliminado: Homer et al., 2015

Eliminado:)

Burned area maps were generated using BAMS (Bastarrika et al., 2014). The Normalized Burn Ratio (NBR), Mid-infrared Burned Index (MIRBI), Global Environmental Monitoring Index (GEMI) and Normalized Difference Vegetation Index (NDVI) were calculated for the pre- and post-fire images and utilized in a supervised classification. The algorithm was trained on manually selected polygons containing (1) clearly burned pixels and (2) spectrally similar but less distinct burned pixels. The algorithm applied a region-growing function between the two types of training polygons, while cut-off values for each variable were extracted from the training polygons. Each classified burned area was then manually edited. When available, the analysts utilized ancillary datasets (e.g. Monitoring Trends in Burn Severity (MTBS, Eidenshink et al., 2007), MODIS active fire points (MOD14 collection 5, Giglio et al., 2009), MODIS burned area (MCD45A1 collection 5, Roy et al., 2008), and aerial imagery) to improve the confidence in their selection of training pixels and manual edits. To maximize the accuracy of the reference dataset, each image pair was classified into burned area extent and visually evaluated and edited independently by three different analysts. A pixel was then classified as burned if it was identified as burned by two of the three analysts. Additional processing details can be found in Vanderhoof et al. (2017).

The CONUS Landsat Burned Area (1988-2013) dataset includes 168 reference data files from 28 Landsat path/rows and six years (1988, 1993, 1998, 2003, 2008, 2013). The temporal length of reference data varies between 16 and 288 days: 37% of pairs of images were separated by less than 50 days, 35% between 50 and 100 days, and 28% by more than 100 days with a maximum time lapse between the pre- and post-fire image of 288 days. The total area of reference data is $5.23 \cdot 10^6$ km^2, of which 0.12% corresponds to burned category, 82.33% to unburned, and 17.55% to unobserved category. Location of reference sites based on TSAs is shown in Fig. 8. With the publication of Hawbaker et al. (2020), this reference dataset is compliant with CEOS-LPVS Stage 4.

2.4.6 NOFFi Greece (2016-2018)

The reference data were obtained using the perimeters produced by the National Observatory of Forest Fires (NOFFi) (http://epadap.web.auth.gr, last access: 25 March 2020) and, specifically, its Object-based Burned Area Mapping (OBAM) service, implemented by the Laboratory of Forest Management and Remote Sensing (FMRS) of the Aristotle University of Thessaloniki. NOFFi-OBAM is an on-demand service, meaning that it is activated after large wildfire events and under explicit requests by the local forest offices. It relies solely on Sentinel-2 imagery and is employed only for fires within Greece. The NOFFi-OBAM algorithm is designed to map fire perimeters and follows a supervised learning approach using a post-fire Sentinel-2 (Level-1C) image, although a pre-fire image is also used for photo-interpretation purposes. The methodology applied to retrieve the fire perimeters is fully described in Tompoulidou et al. (2016). Non-probability sampling design was applied for this dataset; reference sites were selected by convenience based on images previously processed in the NOFFi-OBAM service.

The NOFFi-OBAM fire perimeters were used as the basis for creating the reference data for the NOFFi Greece reference dataset considering the burned area mapping years 2016, 2017 and 2018. For each Sentinel-2 tile ID (e.g. T34SDH) in which fire perimeters were available, the whole time-series of images were visually checked and the date range for the reference file

Eliminado: (post-fire – pre-fire)

Eliminado: Monitoring Trends in Burn Severity (MTBS), MODIS active fire points, MODIS burned area,

Eliminado: Fires with patch sizes of less than 4.05 ha (45 pixels) were removed from the reference data to be comparable with the minimum mapping unit of the BAECV product.

Eliminado:

Eliminado: BAECV

Eliminado: files

Eliminado: 9. The reference dataset's stage of validation should be considered as 3 according to

Eliminado: classification

Eliminado: 7

Eliminado: wildfires

Eliminado: it

Eliminado: 's

Eliminado: s of

Eliminado: there

Eliminado: were available

creation was defined from the first pre-fire image to the last post-fire image. Small fires within the specific time series that were not mapped from the NOFFi-OBAM service were explicitly digitized. Since NOFFi-OBAM only serves Greece, areas outside Greece's official land boundaries (e.g. seas and land areas of neighboring countries) were masked and classified as unobserved surfaces (category = 2). Some burned scars in overlapping border tiles were mapped by using images from those neighboring tiles only if the post-fire image used for the mapping was inside the time span of the former tile ID. For example, the file 'NOFFi_RD_T34SGH_20160710_20160730.shp', includes polygons with preImg/postImg from T35SCK. This can be identified from the preImg, postImg, and tile columns of the file. Clouds and cloud shadows were manually digitized and masked (category = 2), considering the last postImg. Although a non-probability sampling design was applied for this dataset, the NOFFi-OBAM service has been activated for all wildfires greater than 100 ha during the period 2016–2018 and, in many cases, for smaller (or even much smaller) wildfires. Therefore, the dataset contains a representative set of Sentinel-2 tiles that are frequently affected by wildfires in Greece, at least for the given time-period.

The NOFFi Greece dataset includes 34 reference data files from 25 different Sentinel-2 tiles. The temporal length of reference data varies between 5 and 132 days. The total area of reference data is $0.41 \cdot 10^6$ km^2, of which 0.10% corresponds to burned category, 25.83% to unburned, and 74.08% to unobserved category. As shown in Fig. 9, most of the surface of the tiles from this dataset corresponds to sea surface that was labelled as 'no-data' (section 2.2.), this is the reason the unobserved category is so high compared to the rest of the datasets. The location and temporal length of the reference data as well as the number of images used in each reference site are shown Fig. 9. This reference dataset is compliant with CEOS-LPVS Stage 1.

3 Data availability

The BARD compiled in this effort is freely available on the e-cienciaDatos repository (https://doi.org/10.21950/BBQQU7 (Franquesa et al., 2020)). All burned area reference data files have been visually checked, reprojected and reformatted to provide a uniform set of attributes and metadata descriptions to maximize the ease with which these reference files can be used to evaluate global burned area products. A summary of the data included in each dataset is described in Table 2 and 3. Reference shapefiles and metadata files can be downloaded grouped by the datasets described in this publication: FireCCI global (2008), FireCCI global (2003-2014), FireCCI Africa (2016), FireCCI Africa S2 (2016), CONUS Landsat Burned Area (1988-2013), and NOFFi Greece (2016-2018). Plans are underway to expand the Burned Area Reference Database with new reference files that the FireCCI project produces, and we encourage future contributions from the scientific community.

4 Conclusions

BARD is the first publicly available database that compiles and standardizes previously generated validation reference data. Reference datasets included in this database were produced throughout the life of the FireCCI project since 2010, and other initiatives as Landsat Level-3 Science Products and NOFFi projects have joined and contributed to this effort. BARD gathers

Eliminado: out

Eliminado: postfire

Eliminado: formers

Eliminado: 's time span

Eliminado: 20160802

Eliminado: Cloudy images were discarded from the analysis to avoid unmapped areas within the mapped region

Eliminado: 27

Eliminado: files

Con formato: Superíndice

Eliminado: files

Eliminado: is

Eliminado: 10. The reference dataset's stage of validation should be considered as 1 according to the

Eliminado: classification

Eliminado: Burned Area Reference Database

Eliminado: (Franquesa et al., 2020)

Eliminado: BrFLAS Brazil (2015), BAECV

Eliminado: T

Eliminado: will be expanded

Eliminado: are being produced in the FireCCI project and

Eliminado: The Burned Area

and compiles a total of 2661 standardized shapefiles representing reference burned area data generated from approximately 4500 Landsat and Sentinel-2 images and 8 million square kilometres of interpreted land surface. Reference data were produced following the recommendations of the CEOS Calibration/Validation group and visually inspected by two or more experienced interpreters to ensure the accuracy of the data. As BARD is a compilation of datasets that were produced in different projects and years in which different methods were applied (e.g. different sampling methods, sensors, years or region extent), it is highly recommended that the user clearly understands the characteristics of the dataset or datasets that best suits their needs. BA reference database and future updates remedy the lack of an extensive global and regional, multitemporal validation dataset (Humber et al., 2019) and, certainly, can serve as a valuable source for validation of existing products and developing new BA algorithms, particularly those requiring large amounts of training data.

Eliminado: s

Eliminado: come to replace

Eliminado: and forthcoming

5 Appendix A: Supplementary tables

6 Author contributions

MF and EC wrote the first draft of the manuscript. MF has coordinated the manuscript production and prepared the figures, standardized the reference files and organized the BARD database, and managed its publication on the e-cienciaDatos repository. MV provided the CONUS Landsat Burned Area (1988-2013) dataset, DS and IZG provided the NOFFi Greece (2016-2018) dataset. ER provided the FireCCI Africa S2 (2016) dataset, and MP provided the rest of the FireCCI datasets. EC, as the Science Leader of the FireCCI project, managed the overall execution of the project and suggested the preparation of the present article. All the authors have contributed to the writing and reviewing of the manuscript and agreed on the final version.

Eliminado: have written

Eliminado: BAECV

Eliminado: RL, JR and AS provided the BrFLAS Brazil (2015) fire perimeters.

7 Competing interests

The authors declare that they have no conflict of interest.

8 Acknowledgements.

This research has been funded by the FireCCI project (contract no 4000126706/19/I-NB) which is part of the ESA Climate Change Initiative. We thank Joshua J. Picotte (U.S. Geological Survey Earth Resources Observation and Science (EROS) Center, USA), M. Lucrecia Pettinari (University of Alcalá, Spain), Renata Libonati, Julia A. Rodrigues (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), and Alberto W. Setzer (National Institute for Space Research (INPE), Brazil) for their valuable suggestions in the first version of the manuscript. M. Vanderhoof's time was supported by the U.S. Geological

Survey, Land Resources Mission Area, Land Change Science Program. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Alonso-Canas, I., and Chuvieco, E.: Global burned area mapping from ENVISAT-MERIS and MODIS active fire data, Remote Sensing of Environment, 163, 140-152, https://doi.org/10.1016/j.rse.2015.03.011, 2015.

Bastarrika, A., Chuvieco, E., and Martín, M. P.: Mapping burned areas from Landsat TM/ETM+ data with a two-phase algorithm: Balancing omission and commission errors, Remote Sensing of Environment, 115, 1003-1012, https://doi.org/10.1016/j.rse.2010.12.005, 2011.

Bastarrika, A., Alvarado, M., Artano, K., Martinez, M., Mesanza-Moraza, A., Leyre, T., Ramo, R., and Chuvieco, E.: BAMS: A Tool for Supervised Burned Area Mapping Using Landsat Data, Remote Sensing, 6, 12360-12380, https://doi.org/10.3390/rs61212360, 2014.

Boschetti, L., and Roy, D. P.: Defining a fire year for reporting and analysis of global interannual fire variability, Journal of Geophysical Research: Biogeosciences, 113, 2008.

Boschetti, L., Stehman, S. V., and Roy, D. P.: A stratified random sampling design in space and time for regional to global scale burned area product validation, Remote Sensing of Environment, 186, 465-478, https://doi.org/10.1016/j.rse.2016.09.016, 2016.

Boschetti, L., Roy, D. P., Giglio, L., Huang, H., Zubkova, M., and Humber, M. L.: Global validation of the collection 6 MODIS burned area product, Remote Sensing of Environment, 235, 111490, https://doi.org/10.1016/j.rse.2019.111490, 2019.

Working Group on Calibration and Validation - Land Product Validation Subgroup, 2012.

Chuvieco, E., Opazo, S., Sione, W., Del Valle, H., Anaya, J., Di Bella, C., Cruz, I., Manzo, L., López, G., Mari, N., González-Alonso, F., Morelli, F., Setzer, A., Csiszar, I., Kanpandegi, J. A., Bastarrika, A., and Libonati, R.: Global burned-land estimation in Latin America using MODIS composite data, Ecol Appl, 18, 64-79, 10.1890/06-2148.1, 2008.

Chuvieco, E., Lizundia-Loiola, J., Pettinari, M. L., Ramo, R., Padilla, M., Tansey, K., Mouillot, F., Laurent, P., Storm, T., Heil, A., and Plummer, S.: Generation and analysis of a new global burned area product based on MODIS 250 m reflectance bands and thermal anomalies, Earth Syst. Sci. Data, 10, 2015-2031, 10.5194/essd-10-2015-2018, 2018.

Chuvieco, E., Mouillot, F., van der Werf, G. R., San Miguel, J., Tanasse, M., Koutsias, N., García, M., Yebra, M., Padilla, M., Gitas, I., Heil, A., Hawbaker, T. J., and Giglio, L.: Historical background and current developments for mapping burned area from satellite Earth observation, Remote Sensing of Environment, 225, 45-64, https://doi.org/10.1016/j.rse.2019.02.013, 2019.

Cohen, W. B., Yang, Z., and Kennedy, R.: Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync—Tools for calibration and validation, Remote Sensing of Environment, 114, 2911-2924, 2010.

Eidenshink, J. C., Schwind, B., Brewer, K., Zhu, Z.-L., Quayle, B., and Howard, S. M.: A project for monitoring trends in burn severity, Fire Ecology, 3, 3-21, 10.4996/fireecology.0301003, 2007.

Franquesa, M., Vanderhoof, M. K., Stavrakoudis, D., Gitas, I., Roteta, E., Padilla, M., and Chuvieco, E.: BARD: a global and regional validation burned area database, doi:10.21950/BBQQU7, 2020.

Gallego, F. J.: Stratified sampling of satellite images with a systematic grid of points, ISPRS Journal of Photogrammetry and Remote Sensing, 59, 369-376, https://doi.org/10.1016/j.isprsjprs.2005.10.001, 2005.

Giglio, L., Loboda, T., Roy, D. P., Quayle, B., and Justice, C. O.: An active-fire based burned area mapping algorithm for the MODIS sensor, Remote Sensing of Environment, 113, 408-420, https://doi.org/10.1016/j.rse.2008.10.006, 2009.

Giglio, L., Randerson, J. T., van der Werf, G. R., Kasibhatla, P. S., Collatz, G. J., Morton, D. C., and DeFries, R. S.: Assessing variability and long-term trends in burned area by merging multiple satellite fire products, Biogeosciences, 7, 1171-1186, https://doi.org/10.5194/bg-7-1171-2010, 2010.

Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L., and Justice, C. O.: The Collection 6 MODIS burned area mapping algorithm and product, Remote Sensing of Environment, 217, 72-85, https://doi.org/10.1016/j.rse.2018.08.005, 2018.

Grégoire, J. M., Tansey, K., and Silva, J.: The GBA2000 initiative: developing a global burnt area database from SPOT-VEGETATION imagery, Int. J. Remote Sens., 24, 1369–1376, 10.1080/0143116021000044850, 2003.

Hawbaker, T. J., Vanderhoof, M. K., Beal, Y. J., Takacs, J. D., Schmidt, G. L., Falgout, J. T., Williams, B., Fairaux, N. M., Caldwell, M. K., Picotte, J. J., Howard, S. M., Stitt, S., and Dwyer, J. L.: Mapping burned areas using dense time-series of Landsat data, Remote Sensing of Environment, 198, 504-522, 10.1016/j.rse.2017.06.027, 2017.

Hawbaker, T. J., Vanderhoof, M. K., Schmidt, G. L., Beal, Y.-J., Picotte, J. J., Takacs, J. D., Falgout, J. T., and Dwyer, J. L.: The Landsat Burned Area algorithm and products for the conterminous United States, Remote Sensing of Environment, 244, 1-24, https://doi.org/10.1016/j.rse.2020.111801, 2020.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., and Megown, K.: Completion of the 2011 National Land Cover Database for the Conterminous United States - Representing a Decade of Land Cover Change Information, Photogrammetric Engineering and Remote Sensing, 81, 346-354, doi:10.14358/pers.81.5.345, 2015.

Humber, M. L., Boschetti, L., Giglio, L., and Justice, C. O.: Spatial and temporal intercomparison of four global burned area products, Int. J. Digit. Earth, 12, 460-484, 10.1080/17538947.2018.1433727, 2019.

Kennedy, R. E., Yang, Z., and Cohen, W. B.: Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr—Temporal segmentation algorithms, Remote Sensing of Environment, 114, 2897-2910, 2010.

Lizundia-Loiola, J., Otón, G., Ramo, R., and Chuvieco, E.: A spatio-temporal active-fire clustering approach for global burned area mapping at 250 m from MODIS data, Remote Sensing of Environment, 236, 111493, https://doi.org/10.1016/j.rse.2019.111493, 2020.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A.: Good practices for estimating area and assessing accuracy of land change, Remote Sensing of Environment, 148, 42-57, https://doi.org/10.1016/j.rse.2014.02.015, 2014.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., and Kassem, K. R.: Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity, BioScience, 51, 933-938, https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2, 2001.

Padilla, M., Stehman, S. V., and Chuvieco, E.: Validation of the 2008 MODIS-MCD45 global burned area product using stratified random sampling, Remote Sensing of Environment, 144, 187-196, doi:10.1016/j.rse.2014.01.008, 2014.

Padilla, M., Stehman, S. V., Ramo, R., Corti, D., Hantson, S., Oliva, P., Alonso-Canas, I., Bradley, A. V., Tansey, K., Mota, B., Pereira, J. M., and Chuvieco, E.: Comparing the accuracies of remote sensing global burned area products using stratified random sampling and estimation, Remote Sensing of Environment, 160, 114-121, https://doi.org/10.1016/j.rse.2015.01.005, 2015.

Padilla, M., Olofsson, P., Stehman, S. V., Tansey, K., and Chuvieco, E.: Stratification and sample allocation for reference burned area data, Remote Sensing of Environment, 203, 240-255, https://doi.org/10.1016/j.rse.2017.06.041, 2017.

Padilla, M., Wheeler, J., and Tansey, K.: ESA CCI ECV Fire Disturbance: D4.1.1. Product Validation Report, version 2.1. Tech. Rep., https://www.esa-fire-cci.org/sites/default/files/Fire_cci_D4.1.1_PVR_v2.1_0.pdf, 2018.

Plummer, S., Arino, O., Simon, M., and Steffen, W.: Establishing A Earth Observation Product Service For The Terrestrial Carbon Community: The Globcarbon Initiative, Mitigation and Adaptation Strategies for Global Change, 11, 97-111, 10.1007/s11027-006-1012-8, 2006.

Roteta, E., Bastarrika, A., Padilla, M., Storm, T., and Chuvieco, E.: Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa, Remote Sensing of Environment, 222, 1-17, https://doi.org/10.1016/j.rse.2018.12.011, 2019.

Roy, D. P., Frost, P. G. H., Justice, C. O., Landmann, T., Le Roux, J. L., Gumbo, K., Makungwa, S., Dunham, K., Du Toit, R., Mhwandagara, K., Zacarias, A., Tacheba, B., Dube, O. P., Pereira, J. M. C., Mushove, P., Morisette, J. T., Santhana Vannan, S. K., and Davies, D.: The Southern Africa Fire Network (SAFNet) regional burned-area product-validation protocol, Int. J. Remote Sens., 26, 4265-4292, 10.1080/01431160500113096, 2005.

Roy, D. P., Boschetti, L., Justice, C. O., and Ju, J.: The collection 5 MODIS burned area product — Global evaluation by comparison with the MODIS active fire product, Remote Sensing of Environment, 112, 3690-3707, https://doi.org/10.1016/j.rse.2008.05.013, 2008.

Roy, D. P., and Boschetti, L.: Southern Africa validation of the MODIS, L3JRC, and GlobCarbon burned-area products, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, 47, 1032-1044, 2009.

Simon, M., Plummer, S., Fierens, F., Hoelzemann, J. J., and Arino, O.: Burnt area detection at global scale using ATSR-2: The GLOBSCAR products and their qualification, Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2003JD003622, 2004.

Stehman, S. V.: Statistical rigor and practical utility in thematic map accuracy assessment, Photogrammetric Engineering and Remote Sensing, 67, 727-734, 2001.

Stehman, S. V.: Sampling designs for accuracy assessment of land cover, Int. J. Remote Sens., 30, 5243-5272, doi:10.1080/01431160903131000, 2009.

Stehman, S. V., Olofsson, P., Woodcock, C. E., Herold, M., and Friedl, M. A.: A global land-cover validation data set, II: augmenting a stratified sampling design to estimate accuracy by region and land-cover class, Int. J. Remote Sens., 33, 6975-6993, 10.1080/01431161.2012.695092, 2012.

Stehman, S. V., and Foody, G. M.: Key issues in rigorous accuracy assessment of land cover products, Remote Sensing of Environment, 231, 111199, https://doi.org/10.1016/j.rse.2019.05.018, 2019.

Tansey, K., Grégoire, J. M., Stroppiana, D., Sousa, A., Silva, J., Pereira, J. M., Boschetti, L., Maggi, M., Brivio, P. A., and Fraser, R.: Vegetation burning in the year 2000: Global burned area estimates from SPOT VEGETATION data, Journal of Geophysical Research: Atmospheres, 109, 2004.

Tansey, K., Grégoire, J. M., Defourny, P., Leigh, R., Pekel, J. F., Bogaert, E., and Bartholomé, E.: A new, global, multi-annual (2000–2007) burnt area product at 1 km resolution, Geophysical Research Letters, 35, 1-6, 10.1029/2007gl031567, 2008.

Tompoulidou, M., Stefanidou, A., Grigoriadis, D., Dragozi, E., Stavrakoudis, D., and Gitas, I.: The Greek National Observatory of Forest Fires (NOFFi), Fourth International Conference on Remote Sensing and Geoinformation of the Environment, SPIE, 2016.

Vanderhoof, M. K., Fairaux, N., Beal, Y.-J. G., and Hawbaker, T. J.: Validation of the USGS Landsat Burned Area Essential Climate Variable (BAECV) across the conterminous United States, Remote Sensing of Environment, 198, 393-406, https://doi.org/10.1016/j.rse.2017.06.025, 2017.

Vanderhoof, M. K., Fairaux, N. M., Beal, Y.-J. G., and Hawbaker, T. J.: Data Release for the validation of the USGS Landsat Burned Area Product across the conterminous U.S. (ver. 2.0, May 2020): U.S. Geological Survey data release, https://doi.org/10.5066/F7T151VX, 2020.

Zhu, Z., and Woodcock, C. E.: Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change, Remote Sensing of Environment, 152, 217-234, doi:10.1016/j.rse.2014.06.012, 2014.

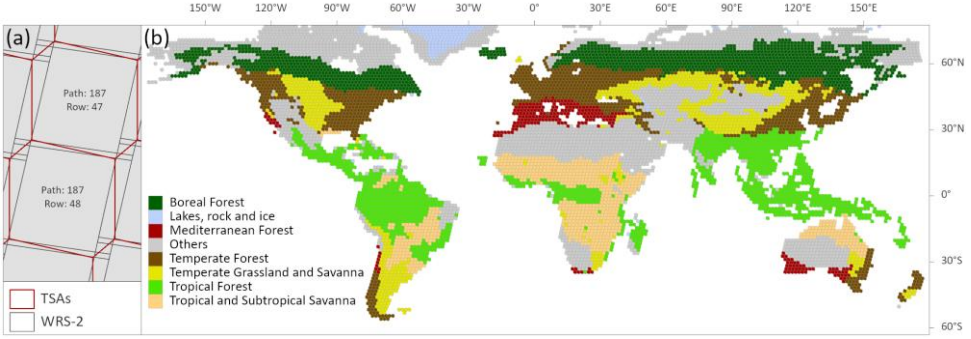

Figure 1: (a)Thiessen scene areas (TSAs) based on Landsat Worldwide Reference System-2 (WRS-2) frames. TSAs are used as non-overlapping spatial units in the sampling design. (b) Distribution of major Olson biomes reclassified as in Padilla et al. (2014).

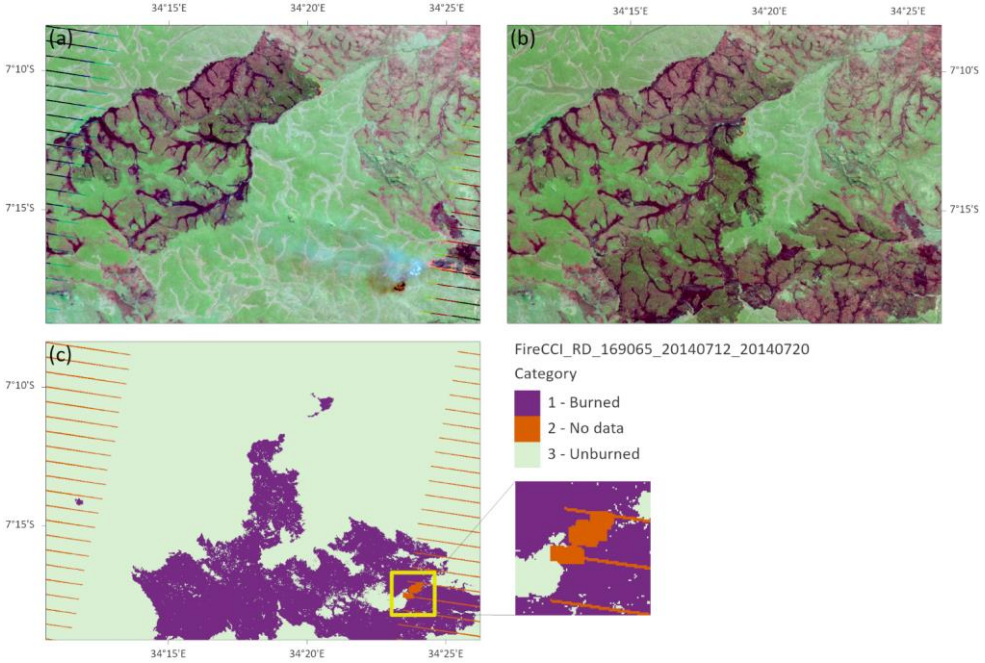

Figure 2: Example of Landsat-7 pre-fire (a) RGB (7,4,3) image and Landsat-8 post-fire (b) RGB (7,5,4) image. Both (a, b), were used to derive the 'FireCCI_RD_169065_20140712_20140720' BA reference file (c) at WRS-2 Landsat 169-065 path-row (East

Eliminado: e

Africa). Time period between both images is 8 days: from 12 June to 20 June 2014. Only the land surface that burns between the two dates is classified as burned, while burned scars in the pre-fire image are assigned to the unburned category. Unobserved pixels on either pre- or post-fire image due to the presence of clouds, cloud-shadows, SLC-gaps or smoke plumes are classified as no-data.

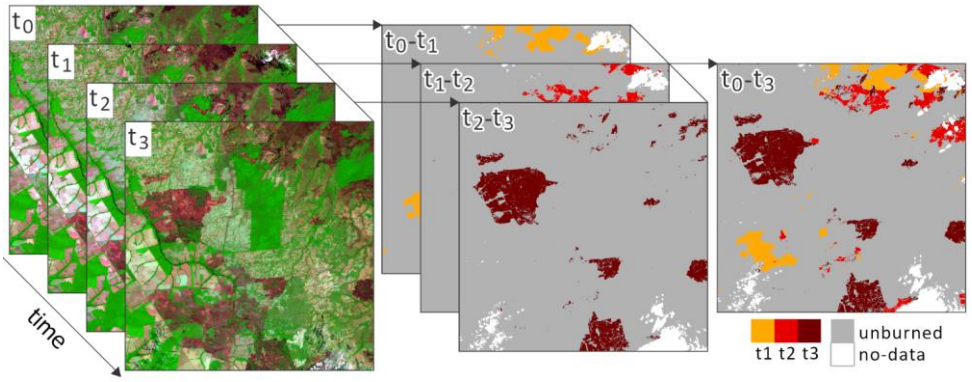

Figure 3: Schematic process of long unit reference data generation. Consecutive image pairs are selected from the multitemporal image series at same location (left: Landsat-8 RGB (7,5,4) images time series) to derive the correspondent short unit reference data files (e.g. Image t_0 and t_1 to obtain the reference data t_0-t_1). From the union of the different short units we generate the long unit reference data (right). The long unit t_0-t_3 includes all the burned scars that occurred between the first image (t_0) and the last image interpreted (t_3), burned scars from the first image (t_0) are not included or mapped. Unobserved areas in any of the images are labeled as no-data in the final long unit reference data. Colours (orange-t1, red-t2, brown-t3) represent the dates in which the burned area patches were observed.

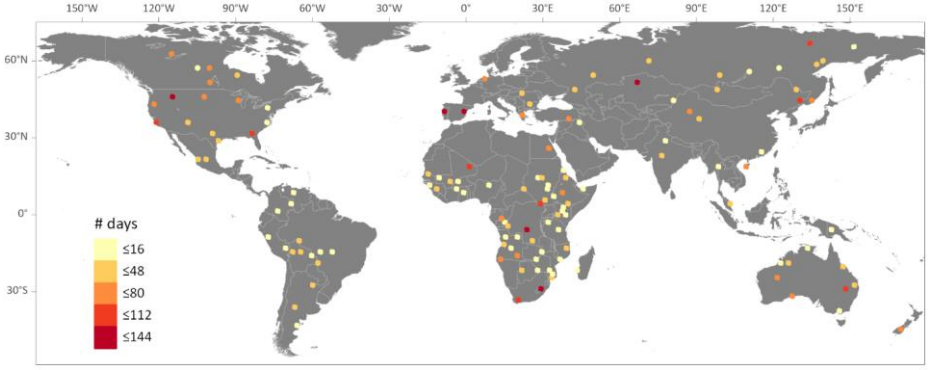

Figure 4: Spatial distribution of the reference sites for FireCCI global (2008) dataset. The legend shows the temporal distance (days) between the pre- and post-fire images used in each validation site for the year 2008.

Eliminado: distance

Eliminado: No

Eliminado: 3

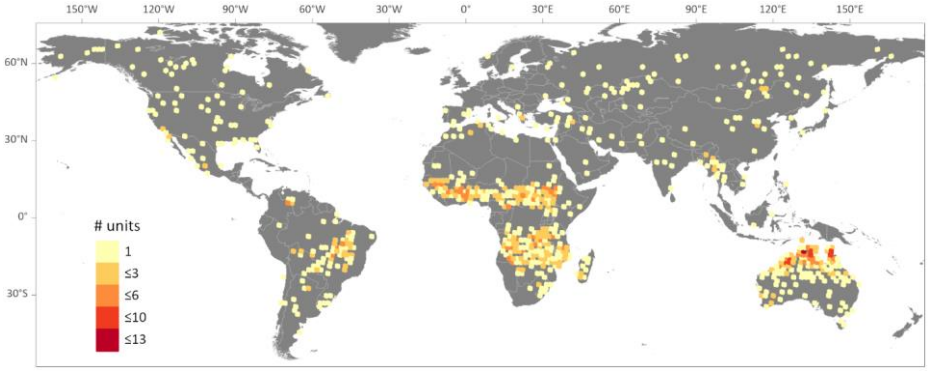

Figure 5: Spatial distribution of the validation Thiessen scene areas (TSAs) for FireCCI global (2003-2014) dataset. The legend shows the total number of reference data files generated for each TSA between the period 2003-2014.

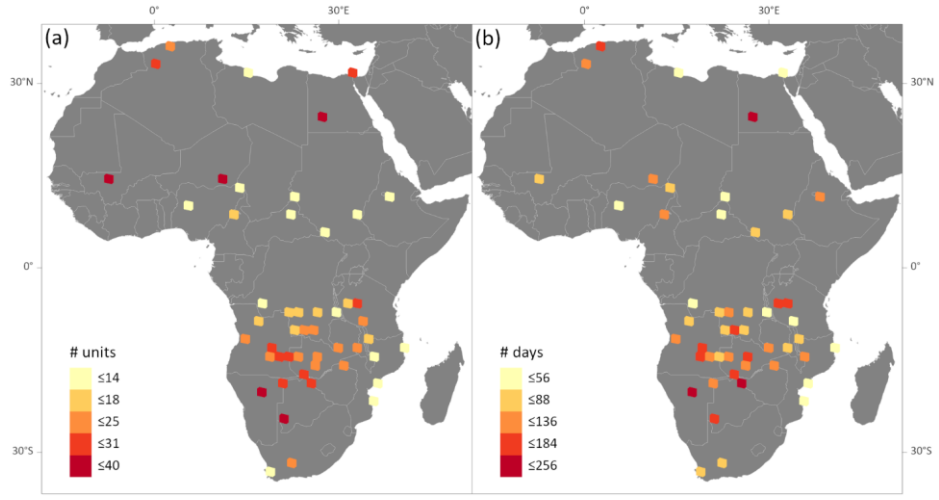

Figure 6: Spatial distribution of the reference sites for the FireCCI Africa (2016) dataset: (a) number of short units interpreted in each validation site and (b) temporal length of the long units.

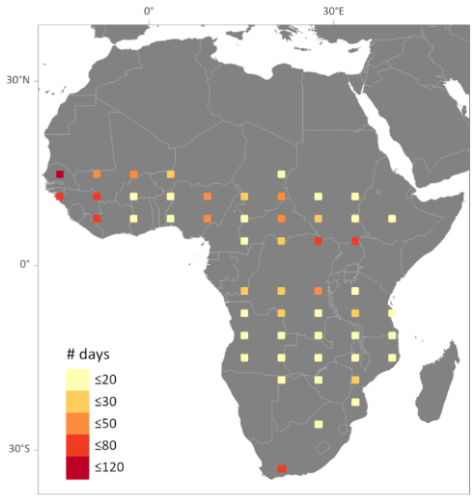

Figure 7: Spatial distribution of the reference sites for FireCCI Africa S2 (2016) dataset. The legend shows the temporal distance (days) between the pre- and post-fire images used in each validation site for the year 2016.

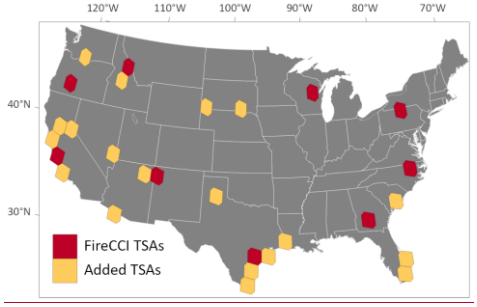

Figure 8: Spatial distribution of the validation Thiessen scene areas (TSAs) for CONUS Landsat Burned Area (1988-2013) dataset. Modified from Vanderhoof et al. (2017). Reference data were generated for each TSA in each of the six sample years (1988, 1993, 1998, 2003, 2008, 2013).

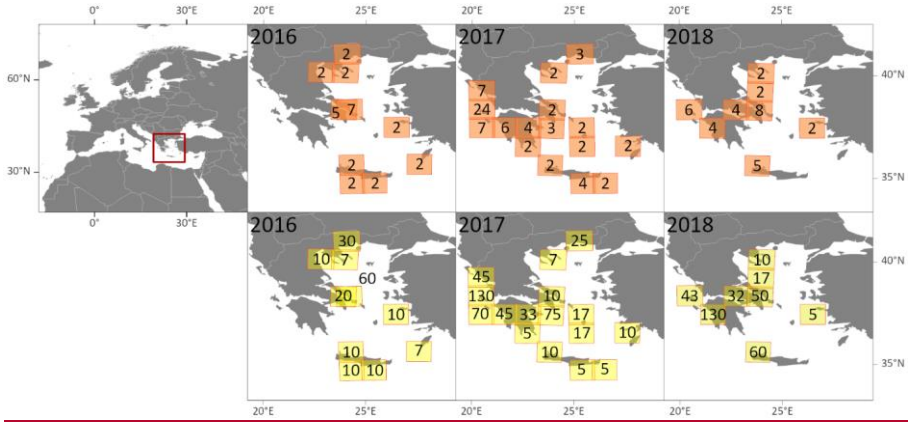

Figure 9: Spatial distribution of validation sites for NOFFi Greece (2016-2018) reference dataset based on Sentinel-2 tiles. The orange figures above show the number of images used in each validation site for each year, whereas the yellow ones below show the temporal length (days) of the reference data files generated in each validation site.

Table 1: Example of the standard attribute table of the reference shapefiles.

Eliminado: 1

category	preDate	postDate	preImg	postImg	path	row	year	area
3	1988-07-05	1988-10-25	LT50150351988187XXX05	LT50150351988299XXX08	15	35	1988	267043.6
2	1988-07-05	1988-10-25	LT50150351988187XXX05	LT50150351988299XXX08	15	35	1988	4557.8
1	1988-07-05	1988-10-25	LT50150351988187XXX05	LT50150351988299XXX08	15	35	1988	2043.3
1	1988-07-05	1988-10-25	LT50150351988187XXX05	LT50150351988299XXX08	15	35	1988	900.4

Table 2: Datasets included in the Burned Area Reference Database. CCI: Climate Change Initiative, CONUS: contiguous United States, NOFFi: National Observatory of Forest Fires, TM: Thematic Mapper, ETM+: Enhanced TM, OLI: Operational Land Imager, CEOS-LPVS: Committee on Earth Observing Satellites-Land Product Validation Subgroup, SRS: Stratified Random Sampling, SS: Systematic Sampling, NPS: Non-probability sampling.

Dataset	Project	Years	Extent	Source Imagery	Sampling Method	CEOS-LPVS Stage	Reference
FireCCI global (2008)	FireCCI	2008	global	Landsat TM, ETM+	SRS	3	Padilla et al. (2014)
FireCCI global (2003-2014)	FireCCI	2003-2014	global	Landsat TM, ETM+, OLI	SRS	3	Padilla et al. (2018)
FireCCI Africa (2016)	FireCCI	2016	Africa	Landsat ETM+, OLI	SRS	3	Padilla et al. (2018)
FireCCI Africa S2 (2016)	FireCCI	2016	Africa	Sentinel-2 MSI	SS	1	Unpublished
CONUS Landsat Burned Area (1988-2013)	Landsat Level-3 Science Products	1988, 1993, 1998, 2003, 2008, 2013	United States	Landsat TM, ETM+, OLI	SRS	4	Vanderhoof et al. (2017;2020)
NOFFi Greece (2016-2018)	NOFFi	2016-2018	Greece	Sentinel-2 MSI	NPS	1	Unpublished

Table 3: Summary of the total area (km^2) of the 3 mapped categories (burned, unburned and no-data) and percentage of each category respect the total area mapped for each dataset. Additionaly, the total land surface and percentage respect the total area interpreted is provided. The region extent and the total number of reference files included in each dataset is also indicated.

Dataset		Region extent	Reference Files (#)	Burned (km^2)	Unburned (km^2)	No-data (km^2)	Land surface (km^2)	Total area (km^2)
FireCCI global (2008)		L5: complete scene L7: central regions without SLC-off gaps	105	23802.26 (1.35%)	1558931.69 (88.35%)	181761.84 (10.30%)	1679627.66 (95.19%)	1764495.79
FireCCI global (2003-2014)		30 x 20 km	1200	27692.96 (3.85%)	516396.61 (71.85%)	174591.03 (24.29%)	674926.47 (93.91%)	718680.59
FireCCI Africa (2016)	SU	30 x 20 km	1052	8398.07 (1.33%)	474349.56 (75.23%)	147821.16 (23.44%)	576181.91 (91.37%)	630568.80
	LU		50	3663.84 (15.72%)	11562.91 (49.61%)	8081.50 (34.67%)	20737.37 (88.97%)	23308.25
FireCCI Africa S2 (2016)		110 x 110 km	52	55583.10 (8.87%)	454013.51 (72.42%)	117317.47 (18.71%)	616483.40 (98.34%)	626914.08
CONUS Landsat Burned Area (1988-2013)		L5-7-8: complete scene	168	6226.45 (0.12%)	4308711 (82.33%)	918382.18 (17.55%)	4251639.569 (81.24%)	5233319.62
NOFFi Greece (2016-2018)		110 x 110 km	34	398.62 (0.10%)	105865.87 (25.83%)	303640.87 (74.08%)	129072.703 (31.49%)	409905.36

Eliminado: BrFLAS: Brazilian Fire-Land-Atmosphere System, BAECV: Burned Area Essential Climate Variable

Table A1: FireCCI global (2008) stratified sampling data. Distribution of sampled (n_h) and total population (N_h) Thiessen scene areas (TSAs) by biome and BA stratum. BA: burned area.

Biome	Number of TSAs sampled (n_h)		Total number of TSAs (N_h)	
	High BA stratum	Low BA stratum	High BA stratum	Low BA stratum
Boreal forest	8	4	215	857
Mediterranean forest	4	3	28	113
Others	3	2	559	2148
Temperate forest	8	9	178	704
Temperate grassland & savanna	4	3	160	637
Tropical forest	9	7	174	696
Tropical & Subtropical savanna	12	29	151	602

Table A2: FireCCI global (2003-2014) stratified sampling data. Distribution of sampled units (n_h) and total population (N_h) by year, biome and BA stratum. H: high, L: Low, BA: burned area.

Biome	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Boreal forest												
Sampled H BA	2	2	2	2	2	2	2	2	2	2	2	2
Sampled L BA	2	2	2	2	2	2	2	2	2	2	2	2
Population H BA	752	745	1344	537	664	826	926	533	1295	1213	726	633
Population L BA	40924	47189	33173	33711	37976	35641	41324	37341	22503	26626	29644	35299
Mediterranean forest												
Sampled H BA	2	2	2	2	2	2	2	2	2	2	2	2
Sampled L BA	2	2	2	2	2	2	2	2	2	2	2	2
Population H BA	179	287	212	292	217	346	329	269	247	314	223	172
Population L BA	8333	7116	7553	7139	7923	6853	7846	7202	7857	5516	7920	8789
Others												
Sampled H BA	2	4	2	6	4	2	2	3	13	2	2	4
Sampled L BA	2	2	2	2	2	2	2	2	2	2	2	2
Population H BA	1694	791	996	768	734	494	798	792	1134	1043	709	764
Population L BA	68577	58049	58971	61564	59484	58978	62512	60303	55806	40999	60530	69961
Temperate forest												
Sampled H BA	2	2	2	2	2	2	2	2	2	2	2	2
Sampled L BA	2	2	2	2	2	2	2	2	2	2	2	2
Population H BA	584	1343	1309	323	951	601	818	1021	907	345	748	729
Population L BA	38622	32424	32747	34122	33850	31544	34438	32708	33925	23146	29994	33036
Temperate grassland & savanna												
Sampled H BA	5	3	4	4	4	6	5	3	3	3	3	5
Sampled L BA	2	2	2	2	2	2	2	2	2	2	2	2
Population H BA	1642	943	1220	996	985	1257	587	858	568	601	488	973

Population L BA	26124	24516	24402	24702	24697	23761	26517	25079	24804	17071	23684	25603
Tropical forest												
Sampled H BA	5	5	5	4	5	3	4	6	3	4	4	4
Sampled L BA	2	2	2	2	2	2	2	2	2	2	2	2
Population H BA	2433	1909	2052	1825	1701	1272	1731	1642	1548	1435	1210	1231
Population L BA	43609	42228	42188	40038	41325	41673	41109	41137	40775	27552	38253	40208
Tropical & subtropical savanna												
Sampled H BA	61	62	55	50	55	60	61	60	50	64	55	62
Sampled L BA	9	8	16	18	14	11	10	10	13	9	10	7
Population H BA	4662	4673	2974	2153	3559	3646	3727	4660	3119	3195	3496	3918
Population L BA	22878	22496	24916	25124	23098	23049	22997	22343	22503	15632	23228	26382

Table A3: FireCCI Africa (2016) stratified sampling data. Distribution of sampled long units and total population by biome and stratum. BA: burned area.

Biome	Number of sampled units (n_h)		Total number of units (N_h)	
	High BA stratum	Low BA stratum	High BA stratum	Low BA stratum
Mediterranean forest	2	2	22	120
Others	2	2	20	549
Temperate grassland & savanna	2	2	24	82
Tropical forest	2	2	96	220
Tropical & subtropical savanna	32	2	393	709

Table A4: CONUS Landsat Burned Area (1988-2013) stratified sampling data. Distribution of sampled and population Thiessen scene areas (TSAs) by biome and stratum. Each sampled TSA was then sampled for 5 separate years; however, high/low BA stratum was determined from 2008, alone. Total number of TSAs is calculated for the contiguous United States (CONUS). BA: burned area.

Biome	Number of TSAs sampled (n_h)		Total number of TSAs (N_h)	
	High BA stratum	Low BA stratum	High BA stratum	Low BA stratum
Temperate forest	6	5	45	179
Mediterranean forest	2	1	2	10
Temperate grassland & savanna	2	3	25	99
Tropical & subtropical savanna	2	2	2	5
Xeric/desert shrub	3	2	17	66