## ESA Climate Change Initiative – Fire_cci
**D4.1.1 Product Validation Report (PVR)**

<table>
<thead>
<tr>
<th><strong>Project Name</strong></th>
<th>ECV Fire Disturbance: Fire_cci Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contract Nº</strong></td>
<td>4000115006/15/I-NB</td>
</tr>
<tr>
<td><strong>Issue Date</strong></td>
<td>12/12/2017</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Author</strong></td>
<td>Marc Padilla, James Wheeler, Kevin Tansey</td>
</tr>
<tr>
<td><strong>Document Ref.</strong></td>
<td>Fire_cci_D4.1.1_PVR_v1.3</td>
</tr>
<tr>
<td><strong>Document type</strong></td>
<td>Public</td>
</tr>
</tbody>
</table>

Project Partners
Prime Contractor/Scientific Lead & Project Management
UAH – University of Alcala (Spain)

Earth Observation Team
UAH – University of Alcala (Spain)
EHU – University of the Basque Country (Spain)
UL – University of Leicester (United Kingdom)
UCL – University College London (United Kingdom)
ISA – School of Agriculture, University of Lisbon (Portugal)

System Engineering
BC – Brockmann Consult (Germany)

Climate Research Group
MPIC – Max Planck Institute for Chemistry (Germany)
IRD - Research Institute for Development (France)
LSCE - Climate and Environmental Sciences Laboratory (France)
VUA - Vrije Universiteit Amsterdam (Netherlands)

Distribution

<table>
<thead>
<tr>
<th>Affiliation</th>
<th>Name</th>
<th>Address</th>
<th>Copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESA</td>
<td>Stephen Plummer (ESA)</td>
<td><a href="mailto:stephen.plummer@esa.int">stephen.plummer@esa.int</a></td>
<td>electronic copy</td>
</tr>
<tr>
<td>Project Team</td>
<td>Emilio Chuvieco (UAH)</td>
<td><a href="mailto:emilio.chuvieco@uah.es">emilio.chuvieco@uah.es</a></td>
<td>electronic copy</td>
</tr>
<tr>
<td></td>
<td>M. Lucrecia Pettinari (UAH)</td>
<td><a href="mailto:mlucrecia.pettinari@uah.es">mlucrecia.pettinari@uah.es</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joshua Lizundia (UAH)</td>
<td><a href="mailto:joshua.lizundia@uah.es">joshua.lizundia@uah.es</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonzalo Otón (UAH)</td>
<td><a href="mailto:gonzalo.oton@uah.es">gonzalo.oton@uah.es</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mihai Tanase (UAH)</td>
<td><a href="mailto:mihai.tanase@uah.es">mihai.tanase@uah.es</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Miguel Ángel Belenguer (UAH)</td>
<td><a href="mailto:miguel.belenguer@uah.es">miguel.belenguer@uah.es</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aitor Bastarrika (EHU)</td>
<td><a href="mailto:aitor.bastarrika@ehu.es">aitor.bastarrika@ehu.es</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ekhi Roteta (EHU)</td>
<td><a href="mailto:ekhi.roteta@gmail.com">ekhi.roteta@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kevin Tansey (UL)</td>
<td><a href="mailto:kjt7@leicester.ac.uk">kjt7@leicester.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marc Padilla Parellada (UL)</td>
<td><a href="mailto:mp489@leicester.ac.uk">mp489@leicester.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>James Wheeler (UL)</td>
<td><a href="mailto:jmw3@leicester.ac.uk">jmw3@leicester.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Philip Lewis (UCL)</td>
<td><a href="mailto:ucfalew@ucl.ac.uk">ucfalew@ucl.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>José Gómez Dans (UCL)</td>
<td><a href="mailto:j.gomez-dans@ucl.ac.uk">j.gomez-dans@ucl.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>James Brennan (UCL)</td>
<td><a href="mailto:james.brennan.11@ucl.ac.uk">james.brennan.11@ucl.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jose Miguel Pereira (ISA)</td>
<td><a href="mailto:jmopereira@gmail.com">jmopereira@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duarte Oom (ISA)</td>
<td><a href="mailto:duarte.oom@gmail.com">duarte.oom@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manuel Campagnolo (ISA)</td>
<td><a href="mailto:mlc@isa.ulisboa.pt">mlc@isa.ulisboa.pt</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thomas Storm (BC)</td>
<td><a href="mailto:thomas.storm@brockmann-consult.de">thomas.storm@brockmann-consult.de</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Johannes Kaiser (MPIC)</td>
<td><a href="mailto:j.kaiser@mpic.de">j.kaiser@mpic.de</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angelika Heil (MPIC)</td>
<td><a href="mailto:a.heil@mpic.de">a.heil@mpic.de</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Florent Mouillot (IRD)</td>
<td><a href="mailto:florent.mouillot@cefe.cnrs.fr">florent.mouillot@cefe.cnrs.fr</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vanesa Moreno (IRD)</td>
<td>mariavanesa.morenodominguez@cefe....</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Philippe Ciais (LSCE)</td>
<td><a href="mailto:philippe.ciais@lsce.ipsl.fr">philippe.ciais@lsce.ipsl.fr</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chao Yue (LSCE)</td>
<td><a href="mailto:chaoyuejoy@gmail.com">chaoyuejoy@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pierre Laurent (LSCE)</td>
<td><a href="mailto:pierre.laurent@lsce.ipsl.fr">pierre.laurent@lsce.ipsl.fr</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guido van der Werf (VUA)</td>
<td><a href="mailto:guido.vander.werf@vu.nl">guido.vander.werf@vu.nl</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ioannis Bistinas (VUA)</td>
<td><a href="mailto:i.bistinas@vu.nl">i.bistinas@vu.nl</a></td>
<td></td>
</tr>
</tbody>
</table>
Summary

This PVR describes the approaches and methods used to assess the quality of BA products coming from the Fire_cci algorithms. The report presents validation results that are representative at global scale and for a multi-year time period.

### Affiliation/Function

<table>
<thead>
<tr>
<th>Prepared</th>
<th>UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Marc Padilla</td>
</tr>
<tr>
<td></td>
<td>James Wheeler</td>
</tr>
<tr>
<td></td>
<td>Kevin Tansey</td>
</tr>
<tr>
<td>Date</td>
<td>04/12/2017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviewed</th>
<th>UAH – Project Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Lucrecia Pettinari</td>
</tr>
<tr>
<td>Date</td>
<td>12/12/2017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Authorized</th>
<th>UAH - Science Leader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Emilio Chuvieco</td>
</tr>
<tr>
<td>Date</td>
<td>12/12/2017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accepted</th>
<th>ESA - Technical Officer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Stephen Plummer</td>
</tr>
<tr>
<td>Date</td>
<td>12/12/2017</td>
</tr>
</tbody>
</table>

This document is not signed. It is provided as an electronic copy.

### Document Status Sheet

<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>31/03/2017</td>
<td>First Issue of the document</td>
</tr>
<tr>
<td>1.1</td>
<td>02/06/2017</td>
<td>Revised document, including answers to the comments on CCI-FIRE-EOPS-MM-17-0043</td>
</tr>
<tr>
<td>1.2</td>
<td>30/09/2017</td>
<td>Updated issue including the SFD validation method.</td>
</tr>
<tr>
<td>1.3</td>
<td>12/12/2017</td>
<td>Revised document, including answers to the comments on CCI-FIRE-EOPS-MM-17-0090</td>
</tr>
</tbody>
</table>

### Document Change Record

<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Request</th>
<th>Location</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>02/06/2017</td>
<td>ESA</td>
<td>Executive summary, Sections 2.3, 3.1.1, Section 2.2, Section 3.4, Annex 6</td>
<td>Small changes in the text, Change applicable documents to reference documents, Added information from the results of the PIR, Figures improved</td>
</tr>
<tr>
<td>1.2</td>
<td>30/09/2017</td>
<td>UL</td>
<td>Executive summary, Sections 2.1, 2.3, 3.2.1, 3.2.2, 5, Section 4.2</td>
<td>Updated text, New section added</td>
</tr>
<tr>
<td>1.3</td>
<td>12/12/2017</td>
<td>ESA</td>
<td>Section 2.2, Section 1, 2.3, 3.1.1, 3.2.2, 5, Section 3.1.2, Section 3.2.1, Section 3.3, Section 6</td>
<td>Moved references to Section 6, Text updated in different paragraphs, Reference to Globcarbon removed, Added information regarding alternative sources of data, Figure 13 title updated, References updated</td>
</tr>
</tbody>
</table>
Table of Contents

1 Executive Summary ........................................................................................................... 7
2 Introduction ......................................................................................................................... 7
   2.1 Purpose of the document .......................................................................................... 7
   2.2 Applicable Documents ............................................................................................ 7
   2.3 Background ................................................................................................................. 8
3 Methods on validation analysis .......................................................................................... 10
   3.1 Reference Data ............................................................................................................ 10
       3.1.1 Reference data generation .................................................................................. 10
       3.1.2 Data structure and naming convention .............................................................. 12
       3.1.3 Metadata ........................................................................................................... 13
   3.2 Sampling design ........................................................................................................... 13
       3.2.1 Sampling units .................................................................................................... 13
       3.2.2 Stratification and sample allocation .................................................................. 16
       3.2.3 Subsample .......................................................................................................... 19
   3.3 Accuracy estimates ....................................................................................................... 19
   3.4 Temporal stability of accuracy .................................................................................. 23
4 Results ............................................................................................................................... 24
   4.1 Global scale ................................................................................................................. 24
   4.2 SFD ............................................................................................................................. 26
5 Discussions and Conclusions ............................................................................................ 27
6 References .......................................................................................................................... 28
Annex 1 Acronyms and abbreviations .................................................................................. 31
Annex 2 README file for preprocess.py ............................................................................. 32
Annex 3 README file for upload.py and classify.py .......................................................... 33
Annex 4 Example of a XML metadata file ............................................................................ 34
Annex 5 Iteration process to allocate sample at year-biome strata on the basis of stratum totals of BA and the $n_{by} \geq 4$ requirement .................................................. 34
Annex 6 Accuracy observations at TSAs ............................................................................ 35

List of Tables

Table 1: Satellite-sensor codes naming convention .......................................................... 12
Table 2: Sampled error matrix on a sampling unit. $e_{ij}$ express the agreements (diagonal cells) or disagreements (off diagonal cells) in terms of area ($m^2$) between the BA product (map) class and the reference class. ......................................................... 19
Table 3: Estimated error matrices and reference burned area ($m^2$) for each product. Standard errors of the estimates are shown in parentheses. ........................................... 24
Table 4: Estimated accuracy of each product. Standard errors of the estimates are shown in parentheses. ................................................................. 24
Table 5: Temporal monotonic trends of accuracy ($b$), ........................................................ 26
**List of Figures**

Figure 1: Example of a Landsat post (19 November 2003) fire image RGB (7, 4, 3) and the derived fire perimeters (yellow lines), at WRS Landsat path-row 97-72 (northeastern Australia)................................................................. 11

Figure 2: Example of attribute table for BA reference data. ....................................... 12

Figure 3: Illustration of short sampling units for a Thiessen scene area (TSA) on a three-dimensional space. Each sampling unit is delimited spatially by a TSA (two-dimensions) and temporally (the third dimension) by the time between two consecutive Landsat images. Images are displayed as false colour composites with SWIR, NIR and red bands in the red, green and blue channels respectively. ....... 14

Figure 4: As in Figure 3 but for the long sampling unit based on consecutive pairs of images.................................................................................................................................................. 14

Figure 5: Spatial distribution of reference data availability for short sampling units. Percentage of time on Thiessen scene areas covered by Landsat image pairs available at the USGS archive separated with 16 days or less between each other, from 2003 to 2014. .......................................................................................................................... 15

Figure 6: Temporal distribution of reference data availability. Monthly percentage of area*time covered by Landsat image pairs separated with 16 days or less between each other.................................................................................................................. 15

Figure 7: Spatial distribution of reference data availability for long sampling units in Africa 2016. Percentage of time on Thiessen scene areas covered by consecutive Landsat image pairs available at the USGS archive separated with 16 days or less between each other covering at least 100 days (sampling units at least 100 days long).......................................................................................................................... 16

Figure 8: Table with the selected BA thresholds $C_{Syb}^*$ for year y and biome b. Grey levels are proportional to threshold values................................................................. 17

Figure 9: Table with the sample sizes $n_h$ for each year y (columns), biome b (rows) and BA level (high BA on the left of the “+” sign and low BA on the right). Grey levels are proportional to the sample size on year and biome strata ($n_{yb}$, the sum of the two $n_h$ of each yb stratum).................................................................................................. 18

Figure 10: Thiessen scene areas (TSAs) with at least one unit selected in the sample and biome stratification based on a reclassification of the 14 Olson biomes (Olson et al. 2001).................................................................................................................. 18

Figure 11: As Figure 10 in but for Africa 2016............................................................... 18

Figure 12: Comparison map between MERIS Fire_cci v4.1 and reference data at sampling unit TSA path 199 and row 52, pre-date 16 January 2010 and post-date 1 February 2010. True burned area is represented in black, true unburned in grey and omission and commission errors in red and green respectively................................. 20

Figure 13: Areas burned between 9 May and 12 July 2016. Colours indicate the range of time when the burn occurred, defined by the Pre- (left panel) and post- (right panel) dates (through blue, green, yellow, orange, red and brown, from 9 May to 12 July 2016). Long sampling unit at TSA path 174 and row 65. Grey represents unburned area and white no-data due to cloud coverage or SLC-OFF problems of ETM+. ................................................................................................................................. 21

Figure 14: Dice of coefficient (DC) and relative bias (relB) for MERIS Fire_cci v4.1 at TSAs. TSAs with reference data but without accuracy measure available are represented by empty polygons (white polygons with grey borders). DC is not
available when there is no BA in the reference data or in the product, and relB is not available when there is no BA in the reference data. ........................................ 25

Figure 15: BA (m²) in the reference data at TSAs. .......................................................... 25

Figure 16: Yearly accuracy estimates. Vertical segments show the 95% confidence intervals. ........................................................................................................................................................................... 26

Figure 17: Estimated accuracy of MCD64 at long sampling units (long su) and at the scale of image pairs (short su). 95% confidence intervals are shown with the error segments. ........................................................................................................................................................................... 27

Figure 18: DC and relB at TSAs for MERIS Fire_cci v4.1. TSAs with reference data but without accuracy measure available are represented by empty polygons (white polygons with grey borders).................................................................................................................................. 35

Figure 19: Same as in Figure 18 but for Ce and Oe. .......................................................... 36

Figure 20: As in Figure 18 but for MCD64........................................................................ 37

Figure 21: As in Figure 20 but for Ce and Oe. .................................................................. 38
1 Executive Summary

The Product Validation Report (PVR) describes the approaches and methods used to assess the quality of burned area (BA) products coming from the Fire_cci algorithms, including those specifically designed for the small fire database (SFD). The current report presents validation results that are representative at global scale and for a multi-year time period.

For a sample of validation sites, BA reference data were generated and compared with BA algorithm outputs, with common temporal interval and spatial coverage. CEOS LPV protocols were used to generate the reference data and peer-reviewed standard methods were used to summarize and express the validation results. Novel methods in BA validation were used to cover a multi-year time period with reference data, using a stratified random sampling of spatio-temporal clusters to optimize the precision of accuracy estimates. The resulting database is novel in BA validation. The temporal stability of BA product estimates is also assessed. A sample was specifically designed for the SFDs, using consecutive images pairs ensuring large temporal overlaps with S-1 and S-2 BA estimates.

The last version of the MERIS Fire_cci v4.1, and additionally the MODIS MCD64 product were validated at global scale for six years through 2005-2011, with a sample of 600 30x20 km spatial windows of pairs of Landsat images separated by 8-16 days. Global accuracy estimates for MERIS Fire_cci v4.1 are similar in terms of magnitude and precision to those observed in MERIS Fire_cci v3.1. No significant temporal trends of accuracy were detected. Globally, commission and omission error ratios were found to be 0.65 (0.05) and 0.81 (0.03) respectively (values within brackets refer to standard errors of accuracy estimates). Those error ratios are comparable to other BA products, which reflect the high complexity of detecting burns. It is important to highlight that part of the errors observed on any product are due to the low spatial resolution of products, in relation to the spatial fragmentation of burns. The same methods will be applied to MODIS Fire_cci v5.0 in the next few months.

Additionally, validation results for a non-Fire_cci product, MCD64, were reported using the new sample specifically designed for the SFDs. Those results are included here as an example of what will be obtained when S-1 and S-2 BA products become available.

2 Introduction

2.1 Purpose of the document

The objective of this Product Validation Report version 2 is to describe and report the validation of MERIS Fire_cci v4.1 and MODIS Fire_cci v5.0 and S-2 Fire_cci SFD v1.0, according to Work Package 4300.

2.2 Applicable Documents

| RD-1 | Fire_cci_PMP_v1.0: Fire_cci Project Management Plan Version 1.0 of Phase 2, issued on 18 November 2015. |
| RD-2 | ECV Fire Disturbance Phase 2 - Proposal prepared for ESA on July 24, 2015 by University of Alcala (UAH, Spain) in association with University of the Basque Country (EHU, Spain) University of Leicester (UL, United Kingdom), University College London (UCL, United Kingdom), School of Agriculture, University of Lisbon (ISA, Portugal), Brockmann Consult |


GmbH (BC, Germany), Max Planck Society for the Advancement of Science represented by Max Planck Institute for Chemistry (MPIC, Germany), Research Institute for Development (IRD, France), Climate and Environmental Sciences Laboratory (LSCE, France), Stichting VU-VUmc (VUA, Netherlands)

2.3 Background

Validation is a critical step of every remote sensing project, as it provides a quantitative assessment of the reliability of results, while facilitating critical information for end users (Congalton and Green 1999). The Committee on Earth Observation Satellites’ Land Product Validation Subgroup (CEOS-LPVS) defines validation as: “The process of assessing, by independent means, the quality of the data products derived from the system outputs”.

CEOS-LPVS defined four stages of validation, based on the coverage and type of reference data sampling (http://lpvs.gsfc.nasa.gov, accessed September 2017):

1. Product accuracy is assessed from a small (typically < 30) set of locations and time periods by comparison with in-situ or other suitable reference data.

2. Product accuracy is estimated over a significant set of locations and time periods by comparison with reference in situ or other suitable reference data. Spatial and temporal consistency of the product and consistency with similar products has been evaluated over globally representative locations and time periods. Results are published in the peer-reviewed literature.

3. Uncertainties in the product and its associated structure are well quantified from comparison with reference in situ or other suitable reference data. Uncertainties are characterized in a statistically rigorous way over multiple locations and time periods representing global conditions. Spatial and temporal consistency of the product and with similar products has been evaluated over globally representative locations and periods. Results are published in the peer-reviewed literature.

4. Validation results for stage 3 are systematically updated when new product versions are released and as the time-series expands.

Through the first decade of the 2000s, BA products were typically subjected to a first stage validation. Globcarbon (Plummer et al. 2007) and L3JRC (Tansey et al. 2008) were validated with independent data derived from 72 Landsat scenes globally distributed mostly from the year 2000. Roy and Boschetti (2009) reported validation results for the MODIS-MCD45 (Roy et al. 2008) product in southern Africa using 11 Landsat scenes, while Chuvieco et al. (2008) validated a regional product for Latin America using 19 Landsat scenes and 9 China–Brazil Earth Resources Satellite (CBERS) scenes. GFED3, which has a coarser spatial resolution of 0.5°, was not formally validated, but some quantification of uncertainty was provided (Giglio et al. 2009; 2010). Recently, the most common BA products were validated with reference data collected by means of probabilistic sampling on a single year, 2008 (Padilla et al. 2014b; Padilla et al. 2015). Later, Boschetti et al. (2016) improved the sampling by specifically including the temporal dimension at the sampling units, but leaving

1 In the context of the CEOS-LPVS guidelines, here uncertainty refers to accuracy obtained from a validation exercise. Commonly uncertainty may be relates to the precision of an estimate.
unsolved the stratification design and sampling allocation to optimally obtain precise accuracy estimates. This was addressed by Padilla et al. (2017) and the main findings were implemented here. The sampling is critical in any validation, to optimize the resources dedicated for the reference data generation. It is particularly critical for the current Fire_cci Phase 2, as validations are intended to be undertaken through several years.

As part of an effort to promote the acceptance of the remote sensing products by external communities, here we provide an independent validation analysis, including the assessment of temporal trends of accuracy. The independence is a critical characteristic of any validation assessment, since it assures that unbiased accuracies are obtained among products. Independence implies that validation datasets are not used during the design of BA algorithms, either for calibration or “tuning” processes. The temporal variability of algorithm performance is one of the key validation aspects to be assessed according to end-user requirements (Heil et al. 2016). The validation then should provide a measure of whether results include temporal trends or not. For the current Fire_cci Phase 2, the reference datasets will be generated to cover at least a decade following a probability sampling, achieving therefore CEOS-LPV validation stage 3.

In-situ reference field data is not feasible to collect or it is very costly. Therefore, many remote sensing validation projects rely on medium spatial (~30 m) resolution images, which are acquired simultaneously as to portray the same ground conditions as the input images from which the validating product is generated. Standard methods on the generation of BA reference data are described in detail by CEOS-LPV (Boschetti et al. 2009).

Accuracy is characterized through cross-tabulation, by accounting for the spatio-temporal coincidences and disagreements on estimates of location and timing of burns between a reference map and the target map. This is the most widely used approach (Boschetti et al. 2016; Padilla et al. 2017; Padilla et al. 2014b; Padilla et al. 2015).

Following the proposal presented to ESA [RD-2] and as stated above, the main objective is to achieve a CEOS-LPV validation stage 3. This implies that the generation of a reference dataset must cover a multi-year time period, initially intended for 2003-2015. The achievement of CEOS-LPV validation stage 4 would then be up to the funding available to support a validation exercise as new product versions are released and as the time-series expands.

In the current version of the PVR, the analyses included all years with available data at the time of writing the report (November 2017). The current version of the PVR (version 1.3) includes the validation of the 2005-2011 MERIS Fire_cci algorithm version 4.1 (hereinafter referred to as MERIS Fire_cci v4.1), the sample over Africa 2016, and validation methods that will be used to validate the small fire database (SFD). Reference data has been generated specifically to be used for the SFD. It was generated from consecutive multiple pairs of Landsat to limit the impact of products’ temporal errors over accuracy inferences. Due to the lower temporal resolution of S-1 and S-2 data compared with that from coarse spatial resolution satellite data, temporal errors of SFDs are expected to be larger. At the time of writing the current report, the Fire_cci SFD BA product data is not yet available for the whole sub-Saharan Africa (target year 2016).
3 Methods on validation analysis

3.1 Reference Data

3.1.1 Reference data generation

This section describes the protocol to generate and document reference information for BA validation. This document is based on the CEOS-CalVal protocol for the validation of burned area products (Boschetti et al. 2009; Padilla et al. 2014a), and has been agreed by the partners involved in the Fire_cci consortium, as well as by the group working on validating MODIS BA information (Boschetti, personal communication).

Reference perimeters are primarily generated from multi-temporal comparison of medium resolution satellite imagery (Landsat-TM or similar), acquired from before and after the fire(s).

After a semi-automatic mapping of burns, a systematic quality control is performed through visual inspection. Each reference dataset is reviewed by a ‘reviewer’ interpreter (M. Padilla) and perimeters with errors are rectified by the ‘author’ interpreter. The review process is done through visual inspection, alternatively displaying the pre- and post-images with the fire perimeters overplayed as yellow lines, and no-data areas as blue non-transparent areas. The reviews are done with the two interpreters (‘author’ and ‘reviewer’) physically at front of the same desktop, to ensure a good and fluid communication and that the improvements needed are clearly understood. This procedure is repeated until no errors are identified.

Based on the experience in Phase 1, the software used to generate reference data, ABAMS, was expected to be found too slow to process the large number of sampling units planned for the current phase. Around 2200 pairs of Landsat images were to be processed for the global sample for 2003-2014 and for the sample specifically designed for the validation of the SFD. That is more than ten times than what was processed in the Fire_cci Phase 1, 200 pairs of images. ABAMS requires the user interaction in two separate times: one for pre-processing of the data and the other for the actual classification. The classification is the most time consuming part. Under a supervised classification, where several classifications might be required until a suitable one is achieved, the time the algorithm needs to do one classification is critical. More importantly, the algorithms of the last versions of ABAMS included large departures from its publication of reference (Bastarrika et al. 2011). The main departure consisted in the removal of the spatial regional algorithm, one of the most important aspects of the original algorithm described in the publication. The remaining algorithm consisted on a classification based on thresholds defined by percentiles observed on training polygons. For these reasons, we decided to implement a standard machine learning algorithm, as described below, embedded in an environment to specifically ingest reference images and produce the reference data with the Fire_cci formats.

The semi-automatic procedure to generate the reference data consists in two steps. In the first step, the pair (pre and post) reflectance satellite images are reformatted to be easily and efficiently used on the second step, the semi-automatic classification of burned/unburned area. The reformatting consists of a co-registration in a region of 30 km width (x) and 20 km high (y) located at the centre of the scene. This is consistent with the sampling design, explained below in Section 3.2.3. The output is a raster file with six bands, with the SWIR, NIR and RED bands of the two Landsat images. Further details can be seen in the documentation (Annex 2) of the Python script where this
reformatting is implemented. This first step is automatic and can be parallelized and be ready well before the interpreter starts with the second step, the semi-automatic classification. For the classification, the interpreter uploads the data in QGIS ([www.qgis.org/](http://www.qgis.org/), accessed September 2017) with pre-defined display settings to digitize the training polygons for burned and unburned areas, and optionally for clouds. The training data is used to fit a Random Forest Classifier (Breiman 2001; Pedregosa et al. 2011), which is a robust classifier used for land cover change detections (Wessels et al. 2016) and increasingly being used in burned area mapping (Ramo and Chuvieco 2017). The classifier takes as input variables the NBR, SWIR and NIR of the pre- and post-dates, and the multitemporal index dNBR. Those spectral regions and indices are specifically useful to discriminate burned areas (Giglio et al. 2009; Goodwin and Collet 2014). Each new actual classification takes about 1 second. The procedure consists in repetitive iterations of visual inspection, delineation of training polygons and classification until no further errors can be perceived on the visual inspection. Optionally, the classification can be overwritten by polygons digitized manually. Once the ‘author’ interpreter is satisfied with the classification, it is then reviewed by the ‘reviewer’ interpreter, which is the same for all reference datasets, and decides whether it is finalized or further rectifications are needed.

The output is an ESRI® shape file with the reference data and metadata as defined below. Further details can be seen in the documentation (Annex 3) of the two Python scripts where this semi-automatic classification is implemented. Figure 1 shows an example of the fire perimeters discrimination.

Parts of the scene that cannot be observed or interpreted, either by clouds or by sensor problems (i.e. SLC-off problems of ETM+) in one of the two images pre or post are classified as no-data. This is to make sure only areas with reliable data are included in the validation process.

![Figure 1: Example of a Landsat post (19 November 2003) fire image RGB (7, 4, 3) and the derived fire perimeters (yellow lines), at WRS Landsat path-row 97-72 (northeastern Australia).](image)
3.1.2 Data structure and naming convention

Each burned area reference file is an ArcGIS™ shape file (.shp), along with the auxiliary files required (.dbf, .prj, shx, .sbn, .xml). The projection is UTM, WGS84, with the UTM zone/row being the zone that is covered by the major part of the scene. The following attribute fields are included in the shape file (Figure 2):

- PreDate. Acquisition date of the image taken before the occurrence of the fire: yyyymmdd (year, month, day).
- PostDate. Acquisition date of the satellite image taken after the fire: yyyymmdd (year, month, day).
- PreImg and PostImg. The pre- and post-fire image names, following this format: satellite-code_Path_Row (e.g. LT5_201_032). The satellite codes are given in Table 1.
- Area (in square metres, m²)
- Category (Observation category):
  - Burned area = 1. This area includes all polygons detected as burned.
  - No-Data = 2. This area 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+)
  - Unburned = 3. This area includes all polygons observed as not burned within the limits of the area covered by the image.

<table>
<thead>
<tr>
<th>Satellite-sensor</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat-4 TM</td>
<td>LT4</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>LT5</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>LE7</td>
</tr>
<tr>
<td>Landsat-8 OLI</td>
<td>LC8</td>
</tr>
</tbody>
</table>

Figure 2: Example of attribute table for BA reference data.
The name of the .shp and associated files is defined as follows:

\[ \text{PRO_RD_YYYYMDD_YYYYMDD_PPPRRR} \]

where:

PRO = Project where the reference data were generated. For the fire perimeters developed within the Fire_cci project, PRO=Fire_cci.

RD = stands for Reference Data

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

pppprrr represents the Landsat Worldwide Reference System (WRS) path and row of the scene (in the case where no Landsat imagery was used, the closest path-row is selected): ppp=path; rrr=row

### 3.1.3 Metadata

The metadata of the reference files is written as an XML document. The metadata contains the author of the reference data file, their institution, the date of creation, the input data sources (names of satellite image files) and the reference of the website of the Fire_cci project. Annex 4 contains an example of a metadata file.

### 3.2 Sampling design

The sampling was designed with two main objectives:

- To provide estimates that can infer accuracy at specific spatial and temporal regions. To achieve that, the dimension of sampling units was defined in terms of spatial and temporal extents, as it is explained in Section 3.2.1.
- To optimally allocate samples through a multi-year time period leading to accuracy estimates as precise as possible. To achieve that, a two-stage cluster sampling allocation was used with optimally defined strata, as explained in Sections 3.2.2 and 3.2.3.

#### 3.2.1 Sampling units

Similarly as in Padilla et al. (2014b; 2015), the spatial dimension of sampling units was based on Landsat WRS-2 to simplify data downloading and processing. The spatial dimension of sampling units was 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. 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).

Following CEOS Validation protocol for BA products (Boschetti et al. 2009), reference data is generated from two consecutive images acquired at the same TSA. Therefore, a sampling unit is delimited spatially by a TSA and temporally by the acquisition dates of consecutive images.

For the global multi-year sample a sampling unit is defined by a pair of images, so the temporality is defined by the acquisition dates of the pair of images, as illustrated in Figure 3. For the sample of Africa 2016 a sampling unit is defined by consecutive pairs of images, so temporally it is defined by the acquisition dates of the first and last
images, as illustrated in ¡Error! No se encuentra el origen de la referencia.. Throughout the document, this sampling unit is referred as “long” unit, as for unit long in time. Contrarily, the unit defined by a pair of images is referred as “short”.

Figure 3: Illustration of short sampling units for a Thiessen scene area (TSA) on a three-dimensional space. Each sampling unit is delimited spatially by a TSA (two-dimensions) and temporally (the third dimension) by the time between two consecutive Landsat images. Images are displayed as false colour composites with SWIR, NIR and red bands in the red, green and blue channels respectively.

Figure 4: As in Figure 3 but for the long sampling unit based on consecutive pairs of images.

The size of a unit $i$, $M_i$, is defined by the multiplication of its size in the spatial dimension (in m²; area of the TSA) and its size in the temporal dimension (in days). Absolute values of $M_i$ size will change if other units were used, however $M_i$ size will remain unchanged in relative terms. A unit $i$ that is twice as large as another in m²-days is also twice as large in km²-seconds, or in any other combination of units. The knowledge of sampling unit sizes is necessary for a later unit subsampling process, and is explained in the sections below. Two consecutive images form a pair whenever they were separated by 16 days or less. It is relevant to limit the time length between two consecutive observations to make sure the spectral signal of a fire that occurred between acquisition times is still present in the latest image.
Landsat imagery with less than 30% of clouds at the USGS archive (http://landsat.usgs.gov/, accessed September 2017) and the temporal requirements between image pairs specified above limited the availability of reference data. Globally from 2003 to 2014 only 26.24% of the area*time is covered by the image pairs available at the USGS archive. In case the ESA archive had Landsat images other than those available at the USGS archive, the amount of available reference data would be larger than that reported here. At the time of designing the sampling the ESA archive did not offer the capability to download large amounts images as we required. Figure 5 shows the spatial distribution of such availability which appears to be affected by cloud global coverage patterns and by Landsat archiving strategies. Figure 6 shows the temporal distribution of reference data availability with clear periodic peaks in the middle of the years and a large increase from 2013 onwards, produced from the start of the Landsat 8 campaign.

Figure 5: Spatial distribution of reference data availability for short sampling units. Percentage of time on Thiessen scene areas covered by Landsat image pairs available at the USGS archive separated with 16 days or less between each other, from 2003 to 2014.

Figure 6: Temporal distribution of reference data availability. Monthly percentage of area*time covered by Landsat image pairs separated with 16 days or less between each other.

Figure 7 shows the spatial distribution of data availability for multiple consecutive pairs of images covering at least 100 consecutive days. This leads to sampling units at least 100 days long. Such a long coverage was set to ensure a good overlap with products generated with S-1 and S-2 imagery, which have much lower temporal resolution compared with those derived from coarser spatial resolution imagery (e.g. MERIS and MODIS).
3.2.2 Stratification and sample allocation

The stratification of sampling units was designed to ensure sufficient sampling in each calendar year, taking into account the major Olson biomes (Olson et al. 2001) and with special focus on regions with high BA. The stratification is based on three levels:

- The first stratification level consisted in assigning each sampling unit to a calendar year. For consistency and simplicity, this assignment was based on the earliest acquisition date of the Landsat image pair. A yearly-stratification level is convenient as it brings flexibility when planning the data collection. Particularly it makes easy to expand the temporal period of study by adding complete years.

- The second stratification level consisted in assigning each sampling unit to the major biome for which the TSA had the maximum area.

- The third stratification level, as in Padilla et al. (2014b; 2015), was based on the BA extent provided by the MCD64 product (Giglio et al. 2009). Sampling units where divided into high and low BA by using a threshold of BA specifically adapted to each year-biome stratum. The sample allocated in each year-biome is proportional to the total BA \( N\overline{BA} \) as recommended by Hansen et al. (1946) for a highly skewed distribution. Padilla et al. (2017) found that an allocation proportional to \( N\sqrt{\overline{BA}} \) lead to more precise accuracy estimates. The study found that, given a same sample size, the use of allocation \( N\overline{BA} \) would lead to standard errors of accuracy measures DC, relB, Ce and Oe (see Section 3.3 for definitions of accuracy measures) around 25%, 50%, 50% and 10% larger respectively, compared with using allocation \( N\sqrt{\overline{BA}} \). That finding came after the samples used in the current report were designed, and a correction was no longer feasible.

Given the available sample size for each year \( y \) and biome \( b \) \( (n_{yb}) \), the threshold was selected to minimize the variance of \( BA_{yb}, V(\overline{BA}_{yb}) \). MCD64 as any other global BA product commonly misses small fires (Hantson et al. 2013; Randerson et al. 2012). If MCD64 misses small fires and they contribute a large area, the allocation method would be less effective. This same shortcoming is described by Hansen et al. (1946) on the survey of business sales, who highlighted that those errors would not introduce bias into the estimates, but would decrease the precision of estimates.
For the global sample for 2003-2014 with short sampling units and using a similar amount of effort in generating reference data as in Fire_cci Phase 1, it was foreseen a sample size of 100 sampling units per year, \( n_y \), at the subsample rate specified later. For a 12-year period, that would amount to 1200 short sampling units. For the sample of Africa for 2016 50 long sampling units were sampled, what lead to approximately 1000 pairs of images (equivalent to the same number of short sampling units). Optimal \( n_{yb} \) was defined with the proportionality of mean BA,

\[
 n_{yb} = n_y \frac{N_{yb} \bar{BA}_{yb}}{N_y \bar{BA}_y}
\]  

At least two sampling units per stratum are needed to compute deviations of BA; hence an iterative process was used (Annex 5) to ensure that all \( n_{yb} \) were \( \geq 4 \) while preserving as much as possible the optimal allocation.

Then, each year-biome (yb) stratum was divided in two parts with an optimal BA threshold. Figure 8 shows the optimal thresholds for each yb stratum, in the scale of the cumulative sum distribution of BA (CS). It ranges from 0 to 1, and it represents the fraction of \( BA_{yb} \) on the sampling units with lower BA than a specific threshold. For example, \( CS_{yb} = 0.5 \) divides a yb in two halves, the one with the sampling units with less BA have the same total BA as the other half. \( CS_{yb} = 0.2 \) makes the half with the sampling units with less BA to have the 20% of \( BA_{yb} \).

<table>
<thead>
<tr>
<th>Year</th>
<th>Tropical Forest</th>
<th>Temperate Forest</th>
<th>Boreal Forest</th>
<th>Tropical and Subtropical savanna</th>
<th>Temperate grassland and savanna</th>
<th>Mediterranean Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>0.16</td>
<td>0.21</td>
<td>0.21</td>
<td>0.13</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>2004</td>
<td>0.21</td>
<td>0.19</td>
<td>0.23</td>
<td>0.12</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>2005</td>
<td>0.11</td>
<td>0.22</td>
<td>0.22</td>
<td>0.26</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>2006</td>
<td>0.15</td>
<td>0.23</td>
<td>0.23</td>
<td>0.28</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>2007</td>
<td>0.24</td>
<td>0.25</td>
<td>0.25</td>
<td>0.28</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>2008</td>
<td>0.17</td>
<td>0.22</td>
<td>0.22</td>
<td>0.20</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>2009</td>
<td>0.13</td>
<td>0.18</td>
<td>0.17</td>
<td>0.16</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>2010</td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
<td>0.16</td>
<td>0.14</td>
<td>0.18</td>
</tr>
<tr>
<td>2011</td>
<td>0.13</td>
<td>0.15</td>
<td>0.13</td>
<td>0.17</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>2012</td>
<td>0.13</td>
<td>0.16</td>
<td>0.17</td>
<td>0.14</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>2013</td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
<td>0.16</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>2014</td>
<td>0.13</td>
<td>0.15</td>
<td>0.13</td>
<td>0.17</td>
<td>0.19</td>
<td>0.18</td>
</tr>
</tbody>
</table>

**Figure 8: Table with the selected BA thresholds \( CS_{yb} \) for year y and biome b. Grey levels are proportional to threshold values.**

The consequent sample sizes \( n_h \) for the global sample 2003-2014 are shown in Figure 9 and the spatial distribution of TSAs with at least one sampling unit selected can be seen in Figure 10. The spatial distribution of TSAs with at least one sampling unit for Africa 2016 is shown in Figure 11. 32 units were allocated in the high BA part of Tropical and Subtropical savanna and two in each of the other strata.
Figure 9: Table with the sample sizes nh for each year y (columns), biome b (rows) and BA level (high BA on the left of the “+” sign and low BA on the right). Grey levels are proportional to the sample size on year and biome strata (nhyb; the sum of the two nh of each yb stratum).

<table>
<thead>
<tr>
<th>Year</th>
<th>Others</th>
<th>Tropical Forest</th>
<th>Temperate Forest</th>
<th>Boreal Forest</th>
<th>Tropical and Subtropical savanna</th>
<th>Temperate grassland and savanna</th>
<th>Mediterranean Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>2+2</td>
<td>4+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2004</td>
<td>5+2</td>
<td>5+2</td>
<td>5+2</td>
<td>5+2</td>
<td>5+2</td>
<td>4+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2005</td>
<td>5+2</td>
<td>5+2</td>
<td>4+2</td>
<td>5+2</td>
<td>5+2</td>
<td>4+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2006</td>
<td>5+2</td>
<td>5+2</td>
<td>4+2</td>
<td>5+2</td>
<td>5+2</td>
<td>4+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2007</td>
<td>4+2</td>
<td>4+2</td>
<td>3+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2008</td>
<td>4+2</td>
<td>4+2</td>
<td>3+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2009</td>
<td>3+2</td>
<td>3+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2010</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2011</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2012</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2013</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td>2014</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
</tbody>
</table>

Figure 10: Thiessen scene areas (TSAs) with at least one unit selected in the sample and biome stratification based on a reclassification of the 14 Olson biomes (Olson et al. 2001).

Figure 11: As Figure 10 in but for Africa 2016.
3.2.3 Subsample

The main advantage of subsampling is that it allows increasing the number of units selected in the first stage. This helps to decrease the variance of accuracy estimates (notice the importance of \( n_b \) as a denominator on Equation 12 in Section 3.3).

Each sampling unit selected was subsampled by a spatial cluster of pixels on a 30 km wide and a 20 km high window located in the geographical centre of the unit. That rectangular size makes it possible to see or almost appreciate single pixels (depending on differences in reflectance between a pixel and its neighbouring area) while the whole image is visualized at a scale of 1:80000 on the screens of 27” used. This hugely reduces the necessity to navigate across the scene in the process of image exploration for the collection of training data and/or revision of image classification. The navigation across a scene is a time consuming task with little or no contribution to reference data generation, thus it is to be avoided as much as possible.

Such subsampling is expected to produce a gain in the estimate precision mainly due to the increase of \( n \) and a within-unit positive correlation (Stehman 1997). The positive correlation implies that pixels within a unit provide similar information, and therefore a sample of them may provide a similar average as the one obtained from all pixels in the unit.

3.3 Accuracy estimates

Commonly in BA validation, accuracy estimates are based on the cross tabulation approach (Congalton and Green 1999; Latifovic and Olthof 2004). The result of the cross tabulation can be represented by the error matrix (Table 2) which expresses the amount of agreements and disagreements in terms of area (m²) between product and reference classifications. A product pixel is coded as “burned” if it was detected as such between the dates defining the temporal dimension of the sampling unit, in the same way as for the reference classification. All other sampled pixels are coded as “unburned” or “no-data”, the latter for unobserved pixels.

**Table 2: Sampled error matrix on a sampling unit.** \( e_{ij} \) express the agreements (diagonal cells) or disagreements (off diagonal cells) in terms of area (m²) between the BA product (map) class and the reference class.

<table>
<thead>
<tr>
<th>Product classification</th>
<th>Reference classification</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burned</td>
<td>Burned ( e_{11} )</td>
<td>( e_{1+} )</td>
</tr>
<tr>
<td></td>
<td>Unburned ( e_{12} )</td>
<td>( e_{2+} )</td>
</tr>
<tr>
<td>Unburned</td>
<td>Burned ( e_{21} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unburned ( e_{22} )</td>
<td></td>
</tr>
<tr>
<td>Col. total</td>
<td>( e_{+1} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( e_{+2} )</td>
<td></td>
</tr>
</tbody>
</table>

The agreement and disagreement areas can be measured in each sampling unit by spatially comparing reference and product binary (burned or unburned) maps. This comparison is performed by overlaying the two vector polygons layers derived from the product and reference datasets. The product binary raster map is converted to polygons and then re-projected to the spatial reference system of the reference dataset. Figure 12 shows an example of a comparison map.
Figure 12: Comparison map between MERIS Fire_cci v4.1 and reference data at sampling unit TSA path 199 and row 52, pre-date 16 January 2010 and post-date 1 February 2010. True burned area is represented in black, true unburned in grey and omission and commission errors in red and green respectively.

Figure 13 illustrates how long sampling units might include areas burned in several time periods; given that reference data it is generated from a temporal series of images. Therefore, the validation of a product can be done at two scales, at the scale of the whole sampling unit, with the binary maps defined by the first and last acquisition dates, and also at the scale of the individual image pairs (similar to the scale of short sampling units), with the binary maps defined by the acquisition dates of the using the series of image pairs. The difference in accuracy estimated from the two scales will give an indication of the effect of the product’s temporal errors and size in time of the sampling unit over the accuracy inferences.

Each cell of the error matrix $e$ of a sampling unit $i$ is defined as its sum across image pair $ps$

$$e = \sum_{ps} e_p$$

(2)

Except for the true unburned area $e_{22}$, which is defined as the area with available data $m$ that is not truly burned, or has commission or omission error through the time series of reference data

$$e_{22} = m - e_{11} - e_{12} - e_{21}$$

(3)
Figure 13: Areas burned between 9 May and 12 July 2016. Colours indicate the range of time when the burn occurred, defined by the Pre- (left panel) and post- (right panel) dates (through blue, green, yellow, orange, red and brown, from 9 May to 12 July 2016). Long sampling unit at TSA path 174 and row 65. Grey represents unburned area and white no-data due to cloud coverage or SLC-OFF problems of ETM+.

Accuracy measures are commonly ratios between combinations of error matrix cells, the commission error ratio,

\[ Ce = \frac{e_{12}}{e_{1+}} \]  

and the omission error ratio,

\[ Oe = \frac{e_{21}}{e_{2+}} \]  

*e ij* refer to the sample values of the error matrix entries. Recent publications (Padilla et al. 2014b; Padilla et al. 2014c; Padilla et al. 2015) used additionally the Dice Coefficient (*DC*) (Dice 1945) and measures of bias. *DC* is particularly useful when comparing product accuracies as it summarizes both error ratios (*Ce* and *Oe*) and expresses the accuracy of the category “burned”. *DC* has a sensible probabilistic interpretation (Dice 1945; Fleiss 1981; Forbes 1995; Hand 1981; Hellden 1980; Liu et al. 2007) as it is the conditional probability that one classifier identifies a pixel as burned, given that the other classifier also identified it as burned (Fleiss 1981).

\[ DC = \frac{2e_{11}}{2e_{11} + e_{12} + e_{21}} \]  

The bias is of interest by end-users (Heil et al. 2016; Mouillot et al. 2014) and can be defined as a total estimate

\[ bias = e_{12} - e_{21} \]  

and in relative terms to the reference BA,

\[ relB = \frac{e_{12} - e_{21}}{e_{1+}} \]  

Global estimates of accuracy are computed taking into account the stratified sampling design and using a stratified combined ratio estimator (Cochran 1977) of the form

\[ \hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{L} N_h \hat{y}_h}{\sum_{h=1}^{L} N_h \hat{x}_h} \]

Where *L* is the number of strata, *N h* is the number of sampling units in stratum *h*, *\( \hat{y}_h \)* and *\( \hat{x}_h \)* are the sample means of *y* i and *x* i at stratum *h*, and *y* i and *x* i are values defined by the
denominator and numerator of the different accuracy measures at sampling unit \( i \). \( y_i \) is defined by \( e_{12}, e_{21}, 2e_{1i}, \) and \( e_{12} - e_{21} \) on \( Ce, Oe, DC \) and \( relB \) respectively. \( x_i \) is defined by \( e_{i+}, e_{+i}, 2e_{i1} + e_{i2} + e_{21} \) and \( e_{+i} \) on \( Ce, Oe, DC \) and \( relB \) respectively.

Because sampling units are of unequal sizes and they are subsampled as explained in Section 3.2.3, the sample means take into account the size of each unit, \( M_i \), and the size of each subsample, \( m_i \)

\[
\bar{y}_h = \frac{1}{n_h} \sum_{i \in h} \frac{M_i y_i}{m_i} \tag{10}
\]
\[
\bar{x}_h = \frac{1}{n_h} \sum_{i \in h} \frac{M_i x_i}{m_i} \tag{11}
\]

where \( n_h \) is the number of sampling units sampled in a stratum. The estimated variance of \( \hat{R} \) is

\[
V(\hat{R}) = \frac{1}{X^2} \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} S_{uh}^2 \tag{12}
\]
\[
S_{uh}^2 = \frac{1}{n_h - 1} \sum_{i \in h} M_i \left( \bar{u}_i - \overline{U}_h \right)^2 \tag{13}
\]

\[
\overline{U}_h = \frac{\sum_{i \in h} u_i}{\sum_{i \in h} M_i} \tag{14}
\]
\[
\bar{u}_i = \frac{u_i}{m_i} \tag{15}
\]
\[
u_i = y_i - Rx_i \tag{16}
\]

Notice that the calculation of the deviation \( S_{uh}^2 \) is based on the means per element (\( \bar{u}_i \) and \( \overline{U}_h \)) and takes into account the size of sampled units (\( M_i \)). With respect to the formulae used in Padilla et al. (2014b; 2015), this is a needed modification to allow for subsampling within each unit. This also represents an improvement as it increases the precision of estimates particularly for units of different sizes (Cochran 1977; Section 9A.1).

Other measurements, such as the bias of the BA in the product and in the reference data (\( BA \) and \( BA_{ref} \) respectively), are expressed as population total estimates of the form

\[
\hat{Y} = \sum_{h=1}^{L} N_h \bar{y}_h \tag{17}
\]

Similarly as above, \( \bar{y}_h \) is the sample mean of \( y_i \), which is defined by \( e_{12} - e_{21}, e_{i+}, \) and \( e_{+i} \) on bias, \( BA \) and \( BA_{ref} \) respectively.

Its variance is

\[
V(\hat{Y}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} S_{yh}^2 \tag{18}
\]
As shown above, reference data is not always available. This presence of no-data is a common source of error in surveys as it may produce bias in the estimates (Cochran 1977; Section 13). The magnitude of such bias depends on the differences between the region with available data and the region without data. In our case, the bias in accuracy estimates depends on how cloud coverage affects the accuracy of BA classifications. For the purposes of the current study, regions with available data were assumed to be similar to those without available data. Thus, the number of sampling units of a stratum \( h \) \( (N_h) \) is defined from the stratum size in terms of area*time \( (M_h) \), assuming that the ratio between number of units with available data and stratum size with available data \( (Na_h/Ma_h) \) is similar to that for the region without data.

\[
N_h = Na_h \frac{M_h}{Ma_h}
\]

### 3.4 Temporal stability of accuracy

Global accuracy estimates are derived for each year, from 2003 to 2014, when product data is available. The objective of the temporal stability assessment is to evaluate the variability of accuracy over time. Following GCOS (2016), the assessment evaluates whether a monotonic trend exists based on the slope \( (b) \) of the relationship between an accuracy measure \( (m) \) and time \( (t) \). Given the small number of observations available (number of years, six), the slope \( b \) of change of accuracy per year is estimated through a nonparametric linear regression (Conover 1999; Section 5.5). For a given accuracy measure \( m \) the slope \( b \) is the median of the slopes between pairs of years \( (b_{ij}) \). For each pair of years \( i \) and \( j \), such than \( i < j \), the “two-year slope” is

\[
b_{ij} = \frac{m_j - m_i}{t_j - t_i}
\]

The temporal monotonic trend of accuracy (i.e. \( b \) different than zero) is tested with the Kendall’s tau \( (\tau) \) statistic (Conover 1999; Section 5.4). A statistically significant test result would indicate that accuracy measure \( m \) presents temporal instability, as it would have a significant increase or decrease over time.

Additionally, accuracy changes can be evaluated particularly for those years where it is expected to find such changes. For example, for products that shift input sensor data, as may be the case for a product that covers a very long time period. Such variations can be directly evaluated by comparing the accuracy inferences between the temporal periods of interest. The current report analyses two products (MERIS Fire_cci v4.1 and MCD64) that use the same input sensor data consistently through all the time period analysed and they are not expected to present accuracy changes at a particular time. The Product Intercomparison Report (Heil et al. 2017) revealed a temporal trend in amount
of data available correlated with burned extents trends. However, the burnt area extent relative to the area that is actually available did not present such temporal trend. This is in agreement with validation results in the current study, that suggests that burns extents and accuracy is similar in areas with and without available data.

4 Results

4.1 Global scale

The population estimates of the error matrix entries \(e_{ij}\) and accuracy measures are presented in Tables 3 and 4. Standard errors of accuracy estimates for the category burn are lower than 0.072 for the two products analysed (MERIS Fire_cci v4.1 and MCD64).

<table>
<thead>
<tr>
<th></th>
<th>(e_{11})</th>
<th>(e_{12})</th>
<th>(e_{21})</th>
<th>(e_{22})</th>
<th>(BA_{\text{ref}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERIS Fire_cci v4.1</td>
<td>1.62e+13</td>
<td>3e+13</td>
<td>7.06e+13</td>
<td>2.8e+16</td>
<td>8.67e+13</td>
</tr>
<tr>
<td></td>
<td>(3e+12)</td>
<td>(7e+12)</td>
<td>(1e+13)</td>
<td>(5e+15)</td>
<td>(1e+13)</td>
</tr>
<tr>
<td>MCD64</td>
<td>2.65e+13</td>
<td>1.43e+13</td>
<td>6.02e+13</td>
<td>2.81e+16</td>
<td>8.68e+13</td>
</tr>
<tr>
<td></td>
<td>(3e+12)</td>
<td>(2e+12)</td>
<td>(1e+13)</td>
<td>(5e+15)</td>
<td>(1e+13)</td>
</tr>
</tbody>
</table>

Table 4: Estimated accuracy of each product. Standard errors of the estimates are shown in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>DC</th>
<th>relB</th>
<th>Ce</th>
<th>Oe</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERIS Fire_cci v4.1</td>
<td>0.243</td>
<td>-0.468</td>
<td>0.650</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.106)</td>
<td>(0.049)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>MCD64</td>
<td>0.416</td>
<td>-0.530</td>
<td>0.350</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.072)</td>
<td>(0.022)</td>
<td>(0.048)</td>
</tr>
</tbody>
</table>

Detailed MERIS Fire_cci v4.1 results of accuracy on TSAs can be seen in Figure 14. BA at the reference data can be seen in Figure 15. TSAs with the highest accuracies (i.e. highest DC) tend to be located where BA is high. Highest accuracies are mainly distributed across the tropical and subtropical savannah of Africa, South America and Australia. On the other hand, BA is underestimated on most TSAs. (i.e. relB<0, represented as red tones in the lower panel of Figure 14). Similar trends can be observed with Ce and Oe, and for MCD64 (Annex 6).
Figure 14: Dice of coefficient (DC) and relative bias (relB) for MERIS Fire_cci v4.1 at TSAs. TSAs with reference data but without accuracy measure available are represented by empty polygons (white polygons with grey borders). DC is not available when there is no BA in the reference data or in the product, and relB is not available when there is no BA in the reference data.

Figure 15: BA (m²) in the reference data at TSAs.

Yearly accuracy estimates can be seen in Figure 16 and results of temporal monotonic trend tests in Table 5. A slight and steady increase in accuracy, although a peak in the
second year (2007), is observed for both products. For MCD64, a slight increase in relB is observed. However, no significant temporal trends were detected.

![Figure 16: Yearly accuracy estimates. Vertical segments show the 95% confidence intervals.](image)

<table>
<thead>
<tr>
<th>Table 5: Temporal monotonic trends of accuracy (b),</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>MERIS Fire_cci v4.1</td>
</tr>
<tr>
<td>MCD64</td>
</tr>
</tbody>
</table>

4.2 SFD

Accuracies of MCD64 at long sampling units were higher than those obtained at the scale of image pairs (“short units”). Bias remained unchanged.
Figure 17: Estimated accuracy of MCD64 at long sampling units (long su) and at the scale of image pairs (short su). 95% confidence intervals are shown with the error segments.

5 Discussions and Conclusions

A new sampling design was developed and tested to estimate the accuracy for a long-time series of BA products. The part of the method that defines the stratification and sampling allocation was recently published (Padilla et al., 2017), and was used as basis to develop a new global validation dataset for Fire_cci Phase 2, including both the global products and the SFD products. A total of 2252 multitemporal pairs of images were processed. To facilitate this processing, a new BA classification code was written adapted to the requirements of the validation files.

MERIS Fire_cci v4.1 and MCD64 were validated and compared using reference BA data generated at 600 sampling units distributed globally through six years (2005-2011). A stratified random sampling of spatio-temporal clusters was used to optimize the precision of accuracy estimates. Temporal stability of accuracy was assessed with available per-year accuracy estimates.

Accuracy levels observed for MERIS Fire_cci 4.1 were similar to those observed in Phase 1 for v3.1 (Padilla and Chuvieco 2014). The very similar or even smaller standard errors of accuracy estimates compared with those from Phase 1 illustrate the efficiency and practicality of the sampling design used here. It is important to take into account that in the current Phase 2 we managed to compute the accuracy for several years, while in Phase 1, global accuracy was available only for one year, 2008.

The distribution of sites with high accuracy is similar to that of sites with high BA in the reference data, mainly over tropical and subtropical savannah, as in results of Phase 1 (Padilla and Chuvieco 2014; Padilla et al. 2014b; Padilla et al. 2015). An explanation of that tendency could be that where BA is high spatio-temporal compactness of burn patches can be higher and then they can be more easily detected by classification algorithms.

The lack of statistically significant temporal trends suggest that the observed slight trends can be just caused by chance. Therefore, it cannot be asserted that such temporal trend exists. The common evolution of DC for both products suggests that
characteristics of the per-year reference data affect similarly the product outputs in a way that some years are better mapped than others. An explanation of this variability can be, similarly as explained above, related to the compactness of burns captured by the sample. Comparisons of accuracies between pairs of years must be done with caution, as the sample cannot be the same.

The different quantity of missing images through years in the case of MERIS, many more missing in 2010 and 2011 than in the other years (Heil et al. 2017), appeared to have limited impact in the estimated accuracies. That suggests that its impact on the temporal errors observed with current reference data accuracy might be much lower than the variance of accuracy estimates.

At the sample of Africa 2016, the differences of MCD64’s accuracy inferences at long sampling units and at the scale of image pairs reflect how temporal reporting errors (Boschetti et al. 2010) are less likely to be included as errors if using long sampling units (long in the temporal dimension) than if the units were defined by image pairs (short sampling units). It is remarkable that the DC is similar for both long and short units, reflecting how temporal errors of MCD64 have limited impact on the estimated accuracies.

6 References


## Annex 1 Acronyms and abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABAMS</td>
<td>Automatic Burned Area Mapping Software</td>
</tr>
<tr>
<td>BA</td>
<td>Burned Area</td>
</tr>
<tr>
<td>CalVal</td>
<td>Calibration and Validation</td>
</tr>
<tr>
<td>CBERS</td>
<td>China-Brazil Earth Resources Satellite</td>
</tr>
<tr>
<td>CCI</td>
<td>Climate Change Initiative</td>
</tr>
<tr>
<td>Ce</td>
<td>Commission error ratio</td>
</tr>
<tr>
<td>CEOS</td>
<td>Committee on Earth Observation Satellites</td>
</tr>
<tr>
<td>CS</td>
<td>Cumulative Sum distribution</td>
</tr>
<tr>
<td>DC</td>
<td>Dice Coefficient</td>
</tr>
<tr>
<td>dNBR</td>
<td>Difference Normalized Burn Ratio</td>
</tr>
<tr>
<td>ESA</td>
<td>European Space Agency</td>
</tr>
<tr>
<td>ESRI</td>
<td>Environmental Systems Research Institute</td>
</tr>
<tr>
<td>ECV</td>
<td>Essential Climate Variables</td>
</tr>
<tr>
<td>ETM+</td>
<td>Enhanced Thematic Mapper +</td>
</tr>
<tr>
<td>GCOS</td>
<td>Global Climate Observing System</td>
</tr>
<tr>
<td>GFED3</td>
<td>Global Fire Emission Database v.3</td>
</tr>
<tr>
<td>LPVS</td>
<td>Land Product Validation Subgroup of CEOS</td>
</tr>
<tr>
<td>OLI</td>
<td>Operational Land Imager</td>
</tr>
<tr>
<td>MCD45</td>
<td>MODIS Collection 5 Burned Area product using the Roy et al. (2008) algorithm</td>
</tr>
<tr>
<td>MCD64</td>
<td>MODIS Collection 6 Burned Area product using the Giglio et al. (2009) algorithm</td>
</tr>
<tr>
<td>MERIS</td>
<td>Medium Resolution Imaging Spectrometer</td>
</tr>
<tr>
<td>MODIS</td>
<td>Moderate Resolution Imaging Spectroradiometer</td>
</tr>
<tr>
<td>NBR</td>
<td>Normalized Burn Ratio</td>
</tr>
<tr>
<td>NIR</td>
<td>Near InfraRed</td>
</tr>
<tr>
<td>Oe</td>
<td>Omission error ratio</td>
</tr>
<tr>
<td>OLI</td>
<td>Operational Land Imager</td>
</tr>
<tr>
<td>PVR</td>
<td>Product Validation Report</td>
</tr>
<tr>
<td>RD</td>
<td>Reference Document</td>
</tr>
<tr>
<td>relB</td>
<td>Relative bias</td>
</tr>
<tr>
<td>RGB</td>
<td>Red-Green-Blue composite</td>
</tr>
<tr>
<td>S-1</td>
<td>Sentinel 1</td>
</tr>
<tr>
<td>S-2</td>
<td>Sentinel 2</td>
</tr>
<tr>
<td>SLC</td>
<td>Scan Line Corrector</td>
</tr>
<tr>
<td>SFD</td>
<td>Small Fire Database</td>
</tr>
<tr>
<td>SWIR</td>
<td>Short Wave InfraRed</td>
</tr>
<tr>
<td>TM</td>
<td>Thematic Mapper</td>
</tr>
<tr>
<td>TSA</td>
<td>Thiessen Scene Area</td>
</tr>
<tr>
<td>URD</td>
<td>User Requirements Document</td>
</tr>
<tr>
<td>UTM</td>
<td>Universal Transverse Mercator</td>
</tr>
<tr>
<td>WGS84</td>
<td>World Geodetic System 1984</td>
</tr>
<tr>
<td>WRS(-2)</td>
<td>Worldwide Reference System (version 2)</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
</tbody>
</table>
Annex 2 README file for preprocess.py

SOFTWARE

preprocess.py

DESCRIPTION

preprocess.py prepares surface reflectance Landsat images (http://landsat.usgs.gov/CDR_LSR.php, accessed February 2017) to be used by the scripts upload.py and classify.py. preprocess.py co-registers two Landsat images acquired at a same path-row, and generates a new raster file with six bands, with the SWIR, NIR and RED bands of the two Landsat images.

PARAMETERS

fpre: full path name of a Landsat uncompressed file (.tar.gz).
fpost: full path name of a Landsat uncompressed file of an image acquired in a later date than the one specified in fpre.
makesubcluster: a string, "True" or "False" to indicate whether the output is to be limited to a 30 km width and 20 km high spatial region located in the centre of the path-row.
outdir: Output directory.

OUTPUT

A subdirectory named with the Landsat file names containing the six band raster file (SWIR, NIR and RED of fpre and fpost on the first three and on the latter three bands respectively). Two folders named "manual" and "training" with (empty) shapefiles needed by upload.py and classify.py.

REQUIREMENTS

Python 2.7.9 (another version may work well as well)
Python libraries loaded on the first lines of the .py file

# EXAMPLE (in bash)
fpre=/somedirectory/LE70010682003282-SC20150930094734.tar.gz
fpost=/somedirectory/LT50010682003290-SC20150930094856.tar.gz
makesubcluster=True
outdir=/someotherdirectory
python preprocess.py $fpre $fpost $makesubcluster $outdir
Annex 3 README file for upload.py and classify.py

SOFTWARE

upload.py and classify.py allow for a supervised classification of burned area from a pair of reflectance images.

DESCRIPTION

Scripts upload.py and classify.py are designed to be run in QGIS with the ScriptRunner. upload.py uploads and displays the output data of preprocess.py. Landsat images are displayed by the rpre and rpos layers.

The only input parameter that is needed is in upload.py, it is the full path name of the input directory (e.g. "/someotherdirectory/LE70010682003282_LT50010682003290").

IMPORTANT: Use quotation marks needed.

classify.py classifies the area of the images according to the training polygons defined in the layer named "training" using a random forest classifier and explanatory variables NBR, dNBR, SWIR and NIR. The classification is overwritten by the polygons defined (if any) in the layer named "manual".

OUTPUT

Shapefile with the format defined in the Fire Disturbance Project Phase 1 for burned area reference data (see Section 5.2.1.2 of the Phase 1 Product Validation Plan on http://www.esa-fire-cci.org/webfm_send/241). It is displayed in QGIS as a layer named "bamap".

REQUIREMENTS

Linux OS (there is some issue with the path names on Windows)
QGIS version >= 2.0.1
QGIS ScriptRunner pluguin
Python libraries loaded on the first lines of the two .py files

DETAILS

If categories of the manual layer have to be modified
- maybe* categories MUST HAVE NEGATIVE VALUES!!
- only positive values for "Burned", "Unburned" and "No-data"
- no any category with value 0

TIP

Classify and check for errors many times, once every few new training polygons are delineated. IMPORTANT: Save layers before running classify.py.

If metadata is needed on output shapefiles, specify your name and your project in the first lines of classify.py.
Annex 4 Example of a XML metadata file

```xml
<metadata>
  <author>Bashir Adamu</author>
  <institution>University of Leicester</institution>
  <modified>13/12/2016</modified>
  <input_datasource>LT52280792005061; LT52280792005077</input_datasource>
  <online_linkage>www.esa-fire-ccl.org</online_linkage>
</metadata>
```

Annex 5 Iteration process to allocate sample at year-biome strata on the basis of stratum totals of BA and the $n_{yb} \geq 4$ requirement

$n_{yb}$ values are initialized with equation 1 and the iteration process consist on

- At year-biome strata with $n_{yb} < 4$
  - $n_{yb} = 4$ ($n_{yb}$ is forced to be four)
  - $BA_{yb} = 0$ ($BA_{yb}$ is forced to be zero)
- Recalculation of $n_{yb}$ not involved in first step with equation 1 but with the updates of the previous steps
- If any $n_{yb} < 4$, repeat the iteration cycle keeping the updates

The iteration process ends when all $n_{yb} \geq 4$. 
Annex 6 Accuracy observations at TSAs

Figure 18: DC and relB at TSAs for MERIS Fire_cci v4.1. TSAs with reference data but without accuracy measure available are represented by empty polygons (white polygons with grey borders).
Figure 19: Same as in Figure 18 but for Ce and Oe.
Figure 20: As in Figure 18 but for MCD64.
Figure 21: As in Figure 20 but for Ce and Oe.