




ESA Climate Change Initiative – Fire_cci

D4.1.1 Product Validation Report (PVR)

Project Name	ECV Fire Disturbance: Fire_cci Phase 2
Contract N°	4000115006/15/I-NB
Issue Date	21/12/2018
Version	2.1
Author	Marc Padilla, James Wheeler, Kevin Tansey
Document Ref.	Fire_cci_D4.1.1_PVR_v2.1
Document type	Public

To be cited as: M. Padilla, J. Wheeler, K. Tansey (2018) ESA CCI ECV Fire Disturbance: D4.1.1. Product Validation Report, version 2.1. Available at: <https://www.esa-fire-cci.org/documents>

	Fire_cci Product Validation Report	Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
		Issue	2.1	Date	22/12/2018
				Page	2

Project Partners

Prime Contractor/ Scientific Lead & Project Management	UAH – University of Alcalá (Spain)
Earth Observation Team	UAH – University of Alcalá (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

Affiliation	Name	Address	Copies
ESA	Stephen Plummer (ESA)	stephen.plummer@esa.int	electronic copy
Project Team	Emilio Chuvieco (UAH)	emilio.chuvieco@uah.es	electronic copy
	M. Lucrecia Pettinari (UAH)	mlucrecia.pettinari@uah.es	
	Joshua Lizundia (UAH)	joshua.lizundia@uah.es	
	Gonzalo Otón (UAH)	gonzalo.oton@uah.es	
	Mihai Tanase (UAH)	mihai.tanase@uah.es	
	Miguel Ángel Belenguer (UAH)	miguel.belenguer@uah.es	
	Aitor Bastarrika (EHU)	aitor.bastarrika@ehu.es	
	Ekhi Roteta (EHU)	ekhi.roteta@gmail.com	
	Kevin Tansey (UL)	kjt7@leicester.ac.uk	
	Marc Padilla Parellada (UL)	mp489@leicester.ac.uk	
	James Wheeler (UL)	jemw3@leicester.ac.uk	
	Philip Lewis (UCL)	ucfalew@ucl.ac.uk	
	José Gómez Dans (UCL)	j.gomez-dans@ucl.ac.uk	
	James Brennan (UCL)	james.brennan.11@ucl.ac.uk	
	Jose Miguel Pereira (ISA)	jmocpereira@gmail.com	
	Duarte Oom (ISA)	duarte.oom@gmail.com	
	Manuel Campagnolo (ISA)	mlc@isa.ulisboa.pt	
	Thomas Storm (BC)	thomas.storm@brockmann-consult.de	
	Johannes Kaiser (MPIC)	j.kaiser@mpic.de	
	Angelika Heil (MPIC)	a.heil@mpic.de	
	Florent Mouillot (IRD)	florent.mouillot@cefe.cnrs.fr	
	Vanesa Moreno (IRD)	mariavanesa.morenodominguez@cefe....	
	Philippe Ciais (LSCE)	philippe.ciais@lsce.ipsl.fr	
Chao Yue (LSCE)	chaoyuejoy@gmail.com		
Pierre Laurent (LSCE)	pierre.laurent@lsce.ipsl.fr		
Guido van der Werf (VUA)	guido.vander.werf@vu.nl		
Ioannis Bistinas (VUA)	i.bistinas@vu.nl		



Fire_cci
Product Validation Report

Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
Issue	2.1	Date	22/12/2018
		Page	3

Summary

This Product Validation Report (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 and regional scale and for a multi-year time period.

	Affiliation/Function	Name	Date
Prepared	UL	Marc Padilla James Wheeler Kevin Tansey	19/12/2018
Reviewed	UAH – Project Manager	Lucrecia Pettinari	21/12/2018
Authorized	UAH - Science Leader	Emilio Chuvieco	21/12/2018
Accepted	ESA - Technical Officer	Stephen Plummer	21/12/2018

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

Document Status Sheet

Issue	Date	Details
1.0	31/03/2017	First Issue of the document
1.1	02/06/2017	Revised document, including answers to the comments on CCI-FIRE-EOPS-MM-17-0043
1.2	30/09/2017	Updated issue including the SFD validation method.
1.3	12/12/2017	Revised document, including answers to the comments on CCI-FIRE-EOPS-MM-17-0090
2.0	22/10/2018	Updated version including validation of new products
2.1	21/12/2018	Revised document, including answers to the comments on CCI-FIRE-EOPS-MM-18-0198

Document Change Record

Issue	Date	Request	Location	Details
1.1	02/06/2017	ESA	Executive summary, Sections 2.3, 3.1.1 Section 2.2 Section 3.4 Annex 6	Small changes in the text Change applicable documents to reference documents Added information from the results of the PIR Figures improved
1.2	30/09/2017	UL	Executive summary, Sections 2.1, 2.3, 3.2.1, 3.2.2, 5 Section 4.2	Updated text New section added
1.3	12/12/2017	ESA	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	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.
2.0	22/10/2018	UL	All document Sections 1, 2.1 Former section 2.2	Added sections with validation of FireCCI50, FireCCI51, FireCCILT10 and FireCCISFD11. Sections updated. Section deleted.



Issue	Date	Request	Location	Details
			Section 2.2	Small changes in the text, and last paragraphs updated.
			Sections 3.1.1, 3.1.2, 3.2.3, 3.4	Small changes in the text.
			Sections 4.1, 4.2, 5	Sections updated.
2.1	21/12/2018	ESA	Sections 1, 2.2, 3.1.1, 3.2.1	Text updated
		UL	Section 3.2	Small changes in the text
		ESA, UL	Section 3.2.2	Small changes in the text
		ESA	Section 3.3	Figure 12 caption updated
		ESA	Section 5	Small changes in the text
		UL	References	New references included.

Table of Contents

1	Executive Summary	8
2	Introduction.....	8
2.1	Purpose of the document.....	8
2.2	Background.....	8
3	Methods on validation analysis.....	11
3.1	Reference Data.....	11
3.1.1	Reference data generation	11
3.1.2	Data structure and naming convention	13
3.1.3	Metadata	15
3.2	Sampling design.....	15
3.2.1	Sampling units	15
3.2.2	Stratification and sample allocation	18
3.2.3	Subsample.....	21
3.3	Accuracy estimates	21
3.4	Temporal stability of accuracy.....	25
4	Results	26
4.1	Global scale.....	26
4.2	SFD	32
5	Discussions and Conclusions	36
6	References	38
	Annex 1 Acronyms and abbreviations.....	42
	Annex 2 README file for preprocess.py	43
	Annex 3 README file for upload.py and classify.py.....	44
	Annex 4 Example of a XML metadata file	45
	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.....	45
	Annex 6 Population estimates of error matrix entries (e_{ij}) and accuracy measures for the global sample 2003-2014.....	46
	Annex 7 Accuracy observations at TSAs for the global sample 2003-2014	47

Annex 8 Population estimates of error matrix entries (e_{ij}) and accuracy measures for the sample of Africa 2016, at long and short sampling units.....52

Annex 9 Accuracy observations at TSAs for the sample of Africa 2016.....54

List of Tables

Table 1: Satellite-sensor codes naming convention 14

Table 2: Example of attribute table for BA reference data. 14

Table 3: 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. 21

Table 4: Temporal monotonic trends of accuracy (b ; monotonic change of accuracy per year). None of them is significantly different from zero at $\alpha=0.05$ according Kendall's tau test. 32

Table 5: Estimated error matrices and reference burned area (m^2) for each product. Standard errors of the estimates are shown in parentheses. 46

Table 6: Estimated accuracy of each product. Standard errors of the estimates are shown in parentheses. 46

Table 7: Estimated error matrices and reference burned area (m^2) for each product. Standard errors of the estimates are shown in parentheses. 52

Table 8: Estimated accuracy of each product. Standard errors of the estimates are shown in parentheses. 53

List of Figures

Figure 1: Example of a Landsat pre (above; 3 November 2003) and post (below; 19 November 2003) fire RGB (7, 4, 3) images and the derived fire perimeters (yellow lines; same in both images), at WRS Landsat path-row 97-72 (northeastern Australia). 13

Figure 2: 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. 16

Figure 3: As in Figure 2 but for the long sampling unit based on consecutive pairs of images. 16

Figure 4: Spatial distribution of reference data availability for short sampling units. Percentage of time on Thiessen scene areas covered by Landsat TM image pairs available at the USGS archive separated with 16 days or less between each other, from 2003 to 2014. 17

Figure 5: Temporal distribution of reference data availability. Monthly percentage of area*time covered by Landsat TM image pairs separated with 16 days or less between each other. 17

Figure 6: Spatial distribution of reference data availability for long sampling units in Africa 2016. Percentage of time on Thiessen scene areas covered by consecutive Landsat TM 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). Data availability is particularly low in the Tropics..... 18

Figure 7: Table with the selected BA thresholds **CSyb** * for year y and biome b. Grey levels are proportional to threshold values. 19

Figure 8: 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). 20

Figure 9: 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). 20

Figure 10: As Figure 9 in but for Africa 2016. 20

Figure 11: Comparison map between FireCCI41 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. 22

Figure 12: Areas burned between 9 May and 12 July 2016. Colours indicate the burning dates. Given that temporal resolution Landsat TM imagery is of several days, burning dates are indicated by ranges of time. Time ranges begin on dates indicated on the left panel and end as indicated on the right panel. (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+. . 23

Figure 13: Estimated accuracy of each product. 95% confidence intervals are shown with the error segments. 26

Figure 14: Dice Coefficient (DC) and relative bias (relB) for 2003-2014 FireCCILT10 at TSAs. TSAs with reference data but without accuracy measures available are represented by empty polygons (white polygons with grey borders). DC is not available when there is no BA in the reference data nor in the product, and relB is not available when there is no BA in the reference data. 27

Figure 15: As in Figure 14 but for 2005-2011 FireCCI41. 28

Figure 16: As in Figure 14 but for FireCCI50. 29

Figure 17: As in Figure 14 but for FireCCI51. 30

Figure 18: 2003-2014 BA (m^2) in the reference data at TSAs. 30

Figure 19: Yearly accuracy estimates. Vertical segments show the 95% confidence intervals. 31

Figure 20: Estimated product accuracies at long sampling units (long su) and at the scale of image pairs (short su). 95% confidence intervals are shown with the error segments. 33

Figure 21: Dice of coefficient (DC) and relative bias (relB) for S2 FireCCISFD11 at TSAs for long sampling units over the sample of Africa 2016. Units without product data or without accuracy measures 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. 33

Figure 22: As in Figure 21 but for FireCCIS1A10. 34

Figure 23: As in Figure 21 but for FireCCILT10. 34



Fire_cci
Product Validation Report

Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
Issue	2.1	Date	22/12/2018
		Page	7

Figure 24: As in Figure 21 but for FireCCI50..... 34

Figure 25: As in Figure 21 but for FireCCI51..... 35

Figure 26: BA (m²) in the reference data at TSAs over the sample of Africa 2016..... 35

Figure 27: Areas burned at TSA path 170 and row 64 between 13 May and 20 October 2016, in the reference data (upper panels; identified by burning dates ranges, as in Figure 12), according to FireCCISFD11 (lower left) and according to FireCCI51 (lower right). Colours indicate the burn detection times. Grey represents unburned area and white no-data due to cloud coverage or SLC-OFF problems of ETM+. . 36

Figure 28: Ce and Oe at TSAs for FireCCILT10. TSAs with reference data but without accuracy measure available are represented by empty polygons (white polygons with grey borders)..... 47

Figure 29: As in Figure 28 but for FireCCI41..... 48

Figure 30: As in Figure 28 but for FireCCI50..... 49

Figure 31: As in Figure 28 but for FireCCI51..... 50

Figure 32: As in Figure 28 but for MCD64..... 51

Figure 33: Ce and Oe for FireCCISDF11 at TSAs for long sampling units over the sample of Africa 2016. Units without product data or without accuracy measure available are represented by empty polygons (white polygons with grey borders).
..... 54

Figure 34: As in Figure 33 but for FireCCIS1A10..... 54

Figure 35: As in Figure 33 but for FireCCILT10..... 54

Figure 36: As in Figure 33 but for FireCCI50..... 55

Figure 37: As in Figure 33 but for FireCCI51..... 55

Figure 38: As in Figure 33 but for MCD64..... 55

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
					Page	8

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. The current report presents validation results that are representative at global scale for the multi-year time period 2003-2014 and for Africa for year 2016.

For a sample of validation sites, BA reference data were generated from Landsat data and compared with BA algorithm outputs, with common temporal interval and spatial coverage. CEOS LPV protocols were used (Boschetti et al. 2010) to generate the reference data and peer-reviewed standard methods (Padilla et al. 2017) were used to summarize and express the validation results. Novel methods in BA validation were developed to cover a multi-year time period with reference data, using a stratified random sampling of spatio-temporal clusters to maximize the precision of accuracy estimates. A validation sample was specifically designed for the small fire dataset (a burned area product derived from Sentinel-1 and -2 images) using Landsat data. Sampling units were defined with long temporal extents (where the temporal extent is the time period covered by the respective reference data), covering over 100 days, ensuring therefore large temporal overlaps with Sentinel-1 and -2 BA estimates. The resulting dataset are novel in BA validation.

At global scale, the FireCCI41, the FireCCI50, the FireCCI51, the FireCCILT10 products, and additionally the MODIS MCD64 product were validated at global scale from 2003 to 2014, with a sample of 1200 30x20 km spatial windows of pairs of Landsat images separated by 8-16 days (a short temporal extent). FireCCI51, with a Dice Coefficient (DC) of 38.2% and relative bias (relB) of -28.0%, was the most accurate among Fire_cci products. DC values were lower than for the MCD64A1 product (DC 47.8% and relB -41.5%), but it showed better relative bias. The lower DC values of FireCCI51 and 50 products are partly caused by the lower temporal reporting accuracy, as the higher performance at long (in time) sampling units indicates.

The FireCCISFD11 and FireCCI50, FireCCI51, FireCCILT10 and MCD64A1 were validated in Africa using 50 long temporal sampling units from 2016 made by consecutive image pairs (referred here as short sampling units). FireCCISFD11 was clearly the most accurate product at long sampling units (DC 77.0% and relB -9.0%), although one of the least accurate at short sampling units (DC 34.2% and relB -9.0%). FireCCI51 and MCD64A1 had similar accuracies at long sampling units, the former slightly higher.


2 Introduction

2.1 Purpose of the document

The objective of this Product Validation Report version 2.0 is to describe and report the validation of MERIS Fire_cci version 4.1 (FireCCI41), MODIS Fire_cci versions 5.0 (FireCCI50), MODIS Fire_cci version 5.1 (FireCCI51), the AVHRR LTDR Fire_cci version 1.0 (FireCCILT10), the Sentinel-2 Small Fire Dataset Fire_cci v1.1 (FireCCISFD11) and the Sentinel-1 Fire_cci v1.0 for Africa (FireCCIS1A10).

2.2 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

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
			Page		9	


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" (European Space Agency, 2007; Morisette et al. 2006).

CEOS-LPVS defined four stages of validation, based on the coverage and type of reference data sampling (<http://lpvs.gsfc.nasa.gov>, accessed October 2018):

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. Uncertainties¹ 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; this can be referred to as stage 1.5 (i.e. better than stage 1 but not at stage 2). Stage 1 validation results were reported by Roy and Boschetti (2009) 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. 2018; 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 unsolved the stratification design and sampling allocation to optimally obtain precise accuracy estimates, and further did not report on any validation results with any reference data arising from the study. This was addressed by Padilla et al. (2017) and the main findings were implemented here. The sampling is critical in any validation, to make the most of the resources dedicated to generate reference data. It is particularly critical for the current Fire_cci Phase 2, as validation is intended to cover several years.

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.

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
			Page		10	

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 were generated to cover twelve years following a probability sampling, achieving therefore CEOS-LPV validation stage 3.

For burned area assessment globally or regionally, the use of in-situ reference field data is not feasible. Therefore, remote sensing validation projects rely on images of medium spatial resolution of around 30 m. Moreover, this spatial resolution corresponds to that used by GCOS (2016) to define end-user requirements on product accuracies.

Reference images 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; Boschetti et al. 2010).

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 (Padilla et al. 2017; Padilla et al. 2014b; Padilla et al. 2015).

The main objective of this validation is to achieve a CEOS-LPV stage 3 validation. This implies that the generation of a reference dataset must cover a multi-year time period. Reference data was generated to cover 12 years, from 2003 to 2014. A CEOS-LPV Stage 4 validation can be achieved using the approach developed here as new product versions are released and as the time series expands.

Additionally, a sample of reference data was specifically generated over Africa 2016 to validate the Small Fire Dataset (SFD). The SFD product is derived from S-1 and, independently, S-2 data. This separate sample uses consecutive images pairs ensuring large temporal overlaps with SFD BA estimates. Due to the lower temporal resolution of the SFDs; observations of the Earth are not normally every day or every other day such as is the case with the global burned area products, , and for this reason temporal errors of the detection date of the SFDs are mitigated through the long temporal reference data extents.

The PVR includes the validation of the MERIS Fire_cci version 4.1 (hereinafter referred to as FireCCI41; available for 2005-2011), the MODIS Fire_cci versions 5.0 and 5.1 (hereinafter referred to as FireCCI50 and FireCCI51 respectively), the AVHRR LTDR Fire_cci v1.0 (hereinafter referred to as FireCCILT10), the Sentinel-2 Small Fire Dataset Fire_cci v1.1 (hereinafter referred to as FireCCISFD11) and the Sentinel-1 Fire_cci v1.0 for Africa (hereinafter referred to as FireCCIS1A10). Additionally, for reference, the MODIS-MCD64A1 Collection 6 (hereinafter referred to as MCD64) was also included. The product of the Copernicus Global Land Service was withdrawn from the analysis as it is already known to have lower accuracy (Padilla et al. 2015) than the MCD64A1 Collection 6 product.

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
					Page	11

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 (Padilla et al. 2014a).

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

After a semi-automatic mapping of burns, a systematic quality control was performed through visual inspection. Each reference dataset was reviewed by a ‘reviewer’ interpreter (M. Padilla) and perimeters with errors were rectified by the ‘author’ interpreter. The review process was done through visual inspection, alternatively displaying the pre- and post-images with the fire perimeters (derived from the semi-automated algorithm) overlain with yellow lines, and no-data areas as blue non-transparent areas. The reviews were 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 was repeated until no visible differences between perimeters and visual inspection were 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 of a classification based on thresholds defined by percentiles observed on training polygons. For these reasons, we decided to use a standard machine learning algorithm, the Random Forest classifier as described below, embedded in a system that ingest the reference images and produce the reference data with the specific Fire_cci formats.

The semi-automatic procedure that was used to generate the reference data consists of 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 TM 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



Fire_cci
Product Validation Report

Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
Issue	2.1	Date	22/12/2018
		Page	12

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/, accessed October 2018) 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 Normalized Burn Ratio (NBR), SWIR and NIR of the pre- and post-dates, and the multitemporal index dNBR (NBR at image acquisition time 2 minus NBR at image acquisition time 1). These spectral regions and indices have been identified as very useful in discriminating burned areas (Giglio et al. 2009; Goodwin and Collet 2014). Each revision of the classification process takes about 1 second. The procedure consists in repetitive iterations of visual inspection, drawing of new training polygons in the software tool (reflecting those burned areas that have not yet been correctly classified, or those incorrectly classified as being burned) 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.

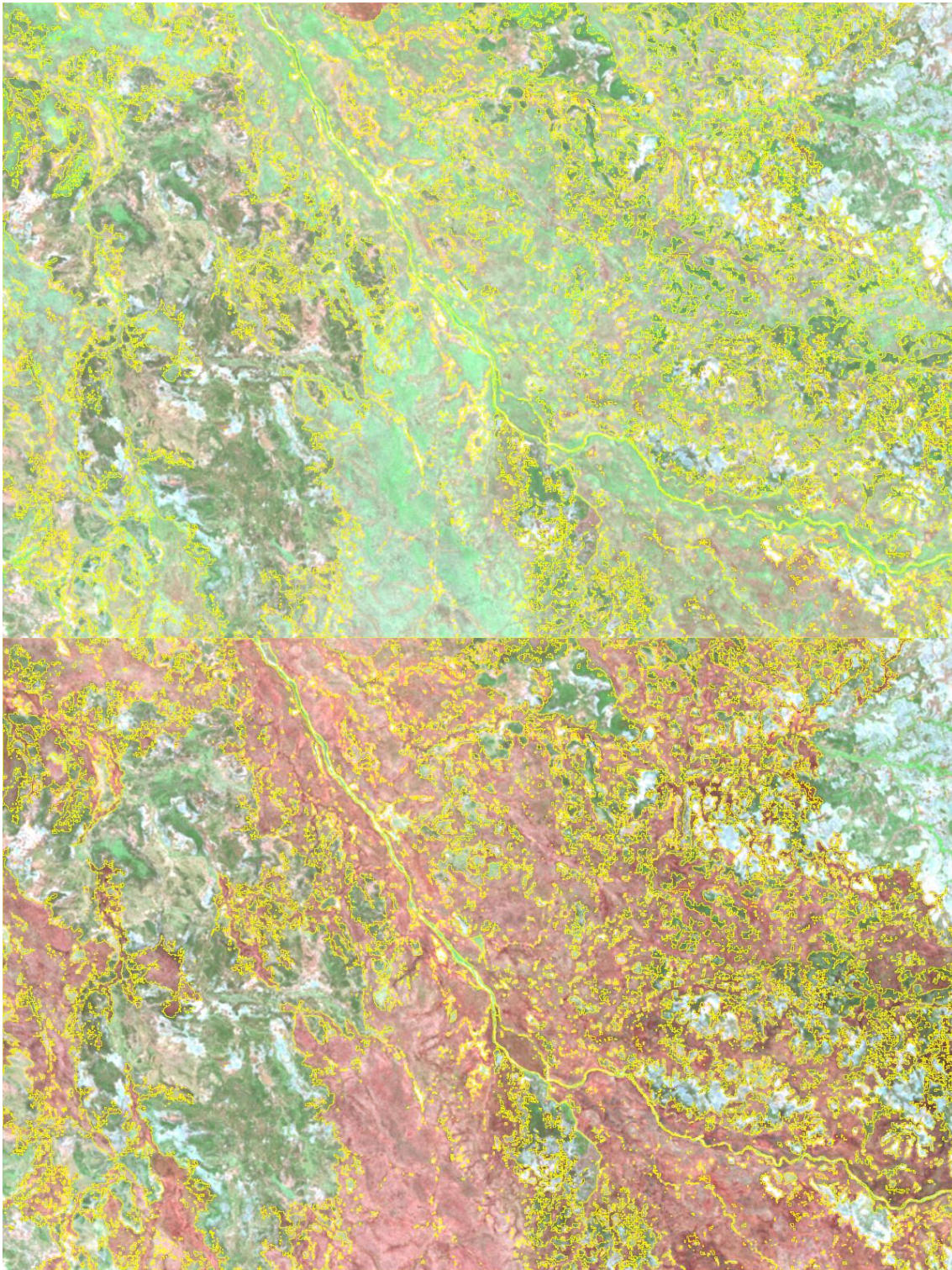


Figure 1: Example of a Landsat pre (above; 3 November 2003) and post (bellow; 19 November 2003) fire RGB (7, 4, 3) images and the derived fire perimeters (yellow lines; same in both images), at WRS Landsat path-row 97-72 (northeastern Australia).

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 (Table 1):

- PreDate. Acquisition date of the image taken before the occurrence of the fire: yyyyymmdd (year, month, day).
- PostDate. Acquisition date of the satellite image taken after the fire: yyyyymmdd (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.

Table 1: Satellite-sensor codes naming convention

Satellite-sensor	Code
Landsat-4 TM	LT4
Landsat-5 TM	LT5
Landsat-7 ETM+	LE7
Landsat-8 OLI	LC8

- 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 2: Example of attribute table for BA reference data.

PreDate	PostDate	Preimg	Postimg	Area	Category
20030630	20030801	LT5_223_066	LT5_223_066	1062000	1
20030630	20030801	LT5_223_066	LT5_223_066	85500	1
20030630	20030801	LT5_223_066	LT5_223_066	933300	1
20030630	20030801	LT5_223_066	LT5_223_066	108000	1
20030630	20030801	LT5_223_066	LT5_223_066	163800	1
20030630	20030801	LT5_223_066	LT5_223_066	1454400	1
20030630	20030801	LT5_223_066	LT5_223_066	38700	1
20030630	20030801	LT5_223_066	LT5_223_066	12600	1
20030630	20030801	LT5_223_066	LT5_223_066	55800	1
20030630	20030801	LT5_223_066	LT5_223_066	244800	1
20030630	20030801	LT5_223_066	LT5_223_066	332100	1


The name of the .shp and associated files is defined as follows:

PRO_RD_YYYYMMDD_YYYYMMDD_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

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
			Page		15	

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.

ppprrr 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 be used to determine accuracy for specific spatial and temporal regions. To achieve this, the dimension of sampling units was defined in terms of spatial and temporal extents, as 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 this, 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

The spatial dimension of sampling units was based on Landsat WRS-2 to simplify data downloading and processing (Padilla et al. 2014b; 2015). 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). 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 2. 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 Figure 3, and nominally every 16 days for Landsat TM. 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 consecutive images is referred as “short”. The assessment of the products is carried out twice, once for the long temporal unit over a spatially limited area (Africa), and second over a short temporal unit for the global products and for a spatially limited area (Africa).

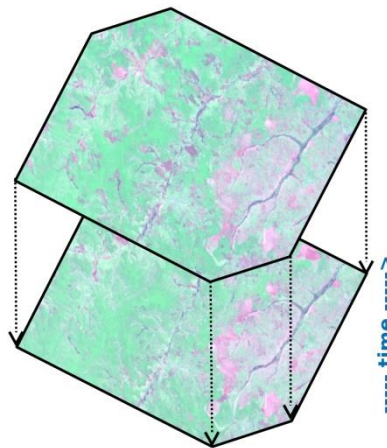


Figure 2: Illustration of short sampling units for a Thiessen scene area (TSA) on a three-dimensional space. Each sampling unit is delimited spatially by a TSA (two-dimensions) and temporally (the third dimension) by the time between two consecutive Landsat images. Images are displayed as false colour composites with SWIR, NIR and red bands in the red, green and blue channels respectively.

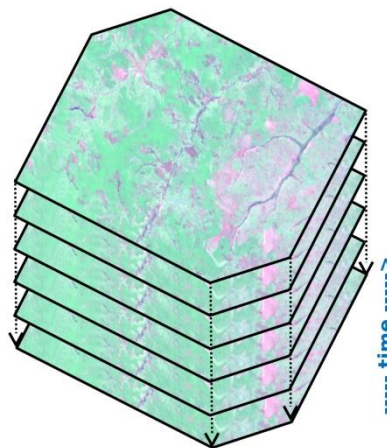


Figure 3: As in Figure 2 but for the long sampling unit based on consecutive pairs of images.

The size of a unit i , M_i , is defined by the multiplication of its size in the spatial dimension (in m^2 ; 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 $m^2 \cdot \text{days}$ is also twice as large in $km^2 \cdot \text{seconds}$, 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. Unlike at the present, at the time of designing the sampling the ESA archive did not offer the capability to download large amounts images as we required. Figure 4 shows the spatial distribution of such availability which appears to be affected by cloud global coverage patterns and by Landsat archiving strategies. Figure 5 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 by the Landsat 8 becoming operational.

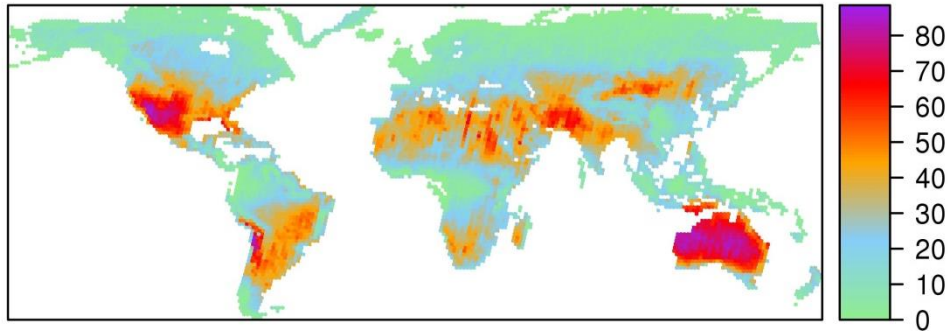


Figure 4: Spatial distribution of reference data availability for short sampling units. Percentage of time on Thiessen scene areas covered by Landsat TM image pairs available at the USGS archive separated with 16 days or less between each other, from 2003 to 2014.

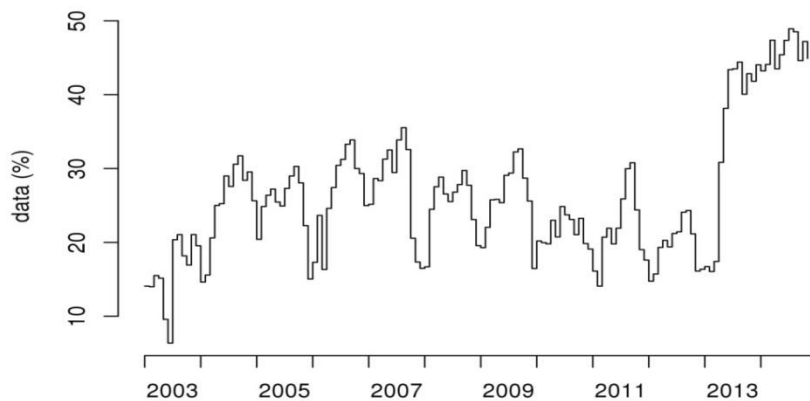


Figure 5: Temporal distribution of reference data availability. Monthly percentage of area*time covered by Landsat TM image pairs separated with 16 days or less between each other.

Figure 6 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 do not observe the surface on a near daily basis.

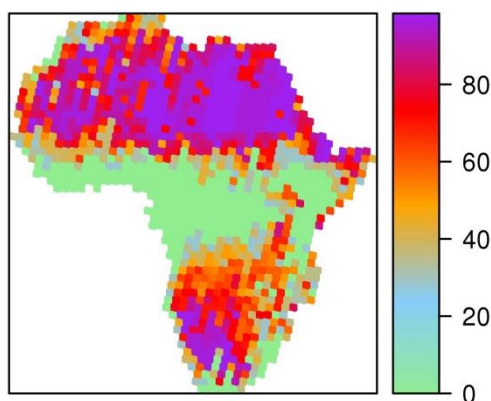


Figure 6: Spatial distribution of reference data availability for long sampling units in Africa 2016. Percentage of time on Thiessen scene areas covered by consecutive Landsat TM image pairs available at the USGS archive separated with 16 days or less between each other covering at least 100 days (sampling units at least 100 days long). Data availability is particularly low in the Tropics.

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 and low fire activity. 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), is based on the BA extent provided by the MODIS-MCD64A1 Collection 5 product (Giglio et al. 2009). Sampling units are 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 ($N\bar{BA}$) as recommended by Hansen et al. (1946) for a highly skewed distribution. Padilla et al. (2017) found that an allocation proportional to $N\sqrt{\bar{BA}}$ lead to more precise accuracy estimates. The study found that, given a same sample size, the use of allocation $N\bar{BA}$ 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 $N\sqrt{\bar{BA}}$.

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} , $V(\bar{BA}_{yb})$. MCD64 as with 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 surveys for 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 of 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 y , 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, which leads 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,

$$n_{yb} = n_y \frac{N_{yb} \overline{BA}_{yb}}{N_y \overline{BA}_y} \quad (1)$$

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 ≥ 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 7 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} .

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Others	0.13	0.12	0.16	0.11	0.15	0.24	0.17	0.13	0.06	0.08	0.14	0.09
Tropical Forest	0.18	0.21	0.2	0.19	0.19	0.25	0.22	0.16	0.21	0.21	0.22	0.25
Temperate Forest	0.21	0.19	0.22	0.32	0.22	0.23	0.23	0.28	0.24	0.29	0.22	0.21
Boreal Forest	0.21	0.14	0.16	0.17	0.18	0.17	0.15	0.19	0.19	0.18	0.13	0.12
Tropical and Subtropical savanna	0.13	0.12	0.22	0.26	0.2	0.16	0.14	0.15	0.21	0.15	0.13	0.1
Temperate grassland and savanna	0.14	0.18	0.15	0.13	0.17	0.13	0.18	0.19	0.18	0.17	0.14	0.13
Mediterranean Forest	0.2	0.23	0.29	0.23	0.24	0.19	0.18	0.22	0.25	0.19	0.18	0.2

Figure 7: 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 8 and the spatial distribution of TSAs with at least one sampling unit selected can be seen in Figure 9. The spatial distribution of TSAs with at least one sampling unit for Africa 2016 is shown in Figure 10. 32 units were allocated in the high BA part of Tropical and Subtropical savanna and two in each of the other strata.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Others	2+2	4+2	2+2	6+2	4+2	2+2	2+2	3+2	13+2	9+2	2+2	4+2
Tropical Forest	5+2	5+2	5+2	4+2	5+2	3+2	4+2	6+2	3+2	4+2	4+2	4+2
Temperate Forest	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2
Boreal Forest	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2
Tropical and Subtropical savanna	61+9	62+8	55+16	50+18	55+14	60+11	61+10	60+10	50+13	55+10	64+9	62+7
Temperate grassland and savanna	5+2	3+2	4+2	4+2	4+2	6+2	5+2	3+2	3+2	4+2	3+2	5+2
Mediterranean Forest	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2	2+2

Figure 8: 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).

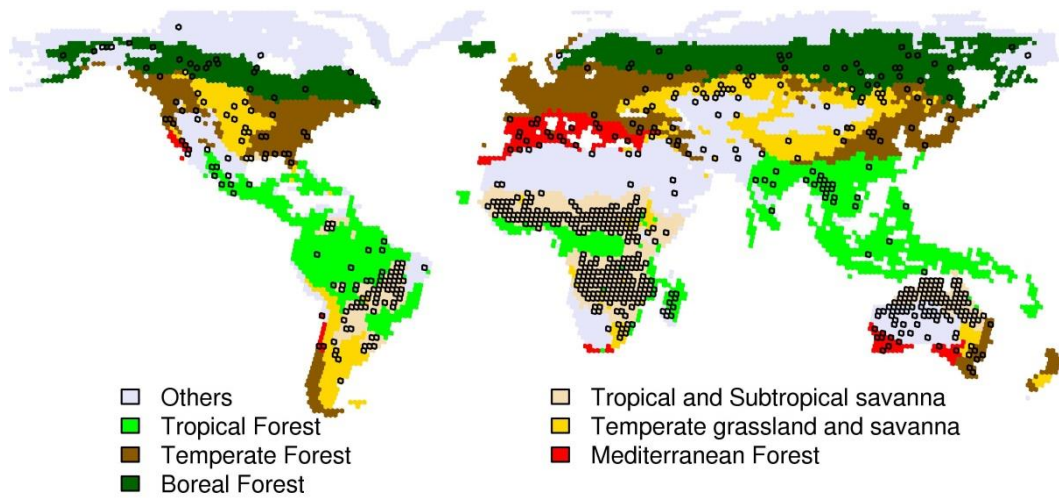


Figure 9: Thiessen scene areas (TSAs) with at least one unit selected in the sample and biome stratification based on a reclassification of the 14 Olson biomes (Olson et al. 2001).

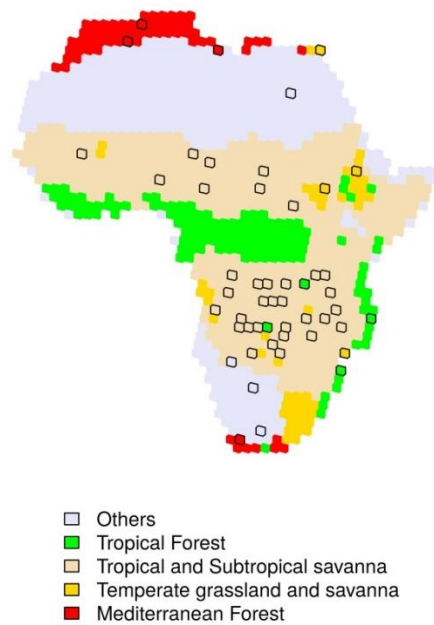


Figure 10: As Figure 9 in but for Africa 2016.

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_h 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 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 which does not actually generate reference data, 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 3) which expresses the amount of agreements and disagreements in terms of area (m^2) 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 3: 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.

Product classification	Reference classification		Row total
	Burned	Unburned	
Burned	e_{11}	e_{12}	e_{1+}
Unburned	e_{21}	e_{22}	e_{2+}
Col. total	e_{+1}	e_{+2}	

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 11 shows an example of a comparison map.

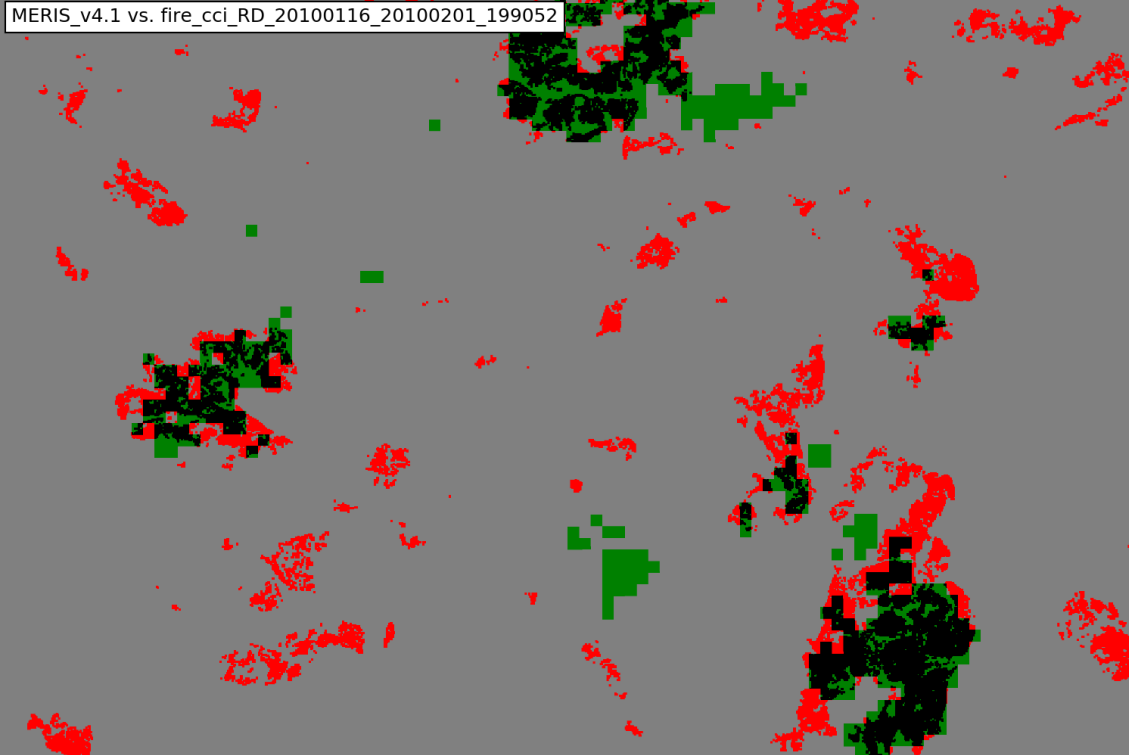


Figure 11: Comparison map between FireCCI41 and reference data at sampling unit TSA path 199 and row 52, pre-date 16 January 2010 and post-date 1 February 2010. True burned area is represented in black, true unburned in grey and omission and commission errors in red and green respectively.

Figure 12 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 series of image pairs used. The difference in accuracy estimated from the two scales will give an indication of the effect of the product's temporal errors 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

$$e = \sum_{p \in i} e_p \quad (2)$$

The exception to this is 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

$$e_{22} = m - e_{11} - e_{12} - e_{21} \quad (3)$$

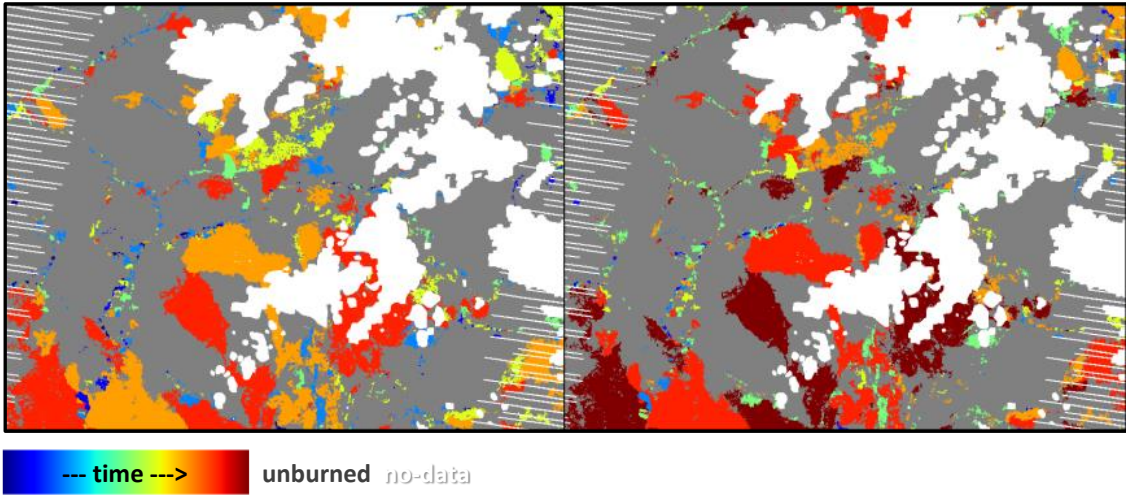


Figure 12: Areas burned between 9 May and 12 July 2016. Colours indicate the burning dates. Given that temporal resolution Landsat TM imagery is of several days, burning dates are indicated by ranges of time. Time ranges begin on dates indicated on the left panel and end as indicated on the right panel. (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,

$$Ce = \frac{e_{12}}{e_{1+}} \quad (4)$$

and the omission error ratio,

$$Oe = \frac{e_{21}}{e_{+1}} \quad (5)$$

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).

$$DC = \frac{2e_{11}}{2e_{11} + e_{12} + e_{21}} \quad (6)$$

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

$$bias = e_{12} - e_{21} \quad (7)$$

and in relative terms to the reference BA,

$$relB = \frac{e_{12} - e_{21}}{e_{+1}} \quad (8)$$

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

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^L N_h \bar{y}_h}{\sum_{h=1}^L N_h \bar{x}_h} \quad (9)$$

Where L is the number of strata, N_h is the number of sampling units in stratum h , \bar{y}_h and \bar{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} , $2e_{11}$ and $e_{12} - e_{21}$ on *Ce*, *Oe*, *DC* and *relB* respectively. x_i is defined by e_{1+} , e_{+1} , $2e_{11} + e_{12} + e_{21}$ and e_{+1} 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

$$\bar{y}_h = \frac{1}{n_h} \sum_{i \in h} \frac{M_i y_i}{m_i} \quad (10)$$

$$\bar{x}_h = \frac{1}{n_h} \sum_{i \in h} \frac{M_i x_i}{m_i} \quad (11)$$

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

$$V(\hat{R}) = \frac{1}{X^2} \sum_{h=1}^L \frac{N_h(N_h - n_h)}{n_h} S_{uh}^2 \quad (12)$$

$$S_{uh}^2 = \frac{1}{n_h - 1} \sum_{i \in h} M_i^2 (\bar{u}_i - \bar{U}_h)^2 \quad (13)$$

$$\hat{U}_h = \frac{\sum_{i \in h} u_i}{\sum_{i \in h} M_i} \quad (14)$$

$$\bar{u}_i = \frac{u_i}{m_i} \quad (15)$$

$$u_i = y_i - R x_i \quad (16)$$

Notice that the calculation of the deviation S_{uh}^2 is based on the means per element (\bar{u}_i and \bar{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 *BAref* respectively), are expressed as population total estimates of the form

$$\hat{Y} = \sum_{h=1}^L N_h \bar{y}_h \quad (17)$$

Similarly as above, \bar{y}_h is the sample mean of y_i , which is defined by $e_{12} - e_{21}$, e_{1+} and e_{+1} on *bias*, *BA* and *BAref* respectively.

Its variance is

$$V(\hat{Y}) = \sum_{h=1}^L \frac{N_h(N_h - n_h)}{n_h} S_{yh}^2 \quad (18)$$

$$S_{yh}^2 = \frac{1}{n_h - 1} \sum_{i \in h} M_i^2 (\bar{y}_i - \bar{Y}_h)^2 \quad (19)$$

$$\hat{\bar{Y}}_h = \frac{\sum_{i \in h} y_i}{\sum_{i \in h} M_i} \quad (20)$$

$$\bar{y}_i = \frac{y_i}{m_i} \quad (21)$$

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.

$$N_h = Na_h \frac{M_h}{Ma_h} \quad (22)$$

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, twelve), 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 that $i < j$, the “two-year slope” is

$$b_{ij} = \frac{m_j - m_i}{t_j - t_i} \quad (23)$$

The temporal monotonic trend of accuracy (i.e. b different than zero) is tested with the Kendall’s tau (τ) 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 seven products and all 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 seems to be in agreement with the assumption mentioned above in the last paragraph of Section 3.3 that product accuracies are similar in areas with and without available data.

4 Results

4.1 Global scale

The population estimates of accuracy measures are presented in Figure 13. Detailed results of error matrix entries (e_{ij}) and accuracy measures are presented as tables in Annex 6. Accuracy results show how MCD64 is the most accurate product, followed by FireCCI51 and FireCCI50, according to the Dice Coefficient and commission error ratio. FireCCI51 had the lowest RelB followed closely by FireCCI50 and MCD64 (differences not statistically significant at the 0.05 confidence level).

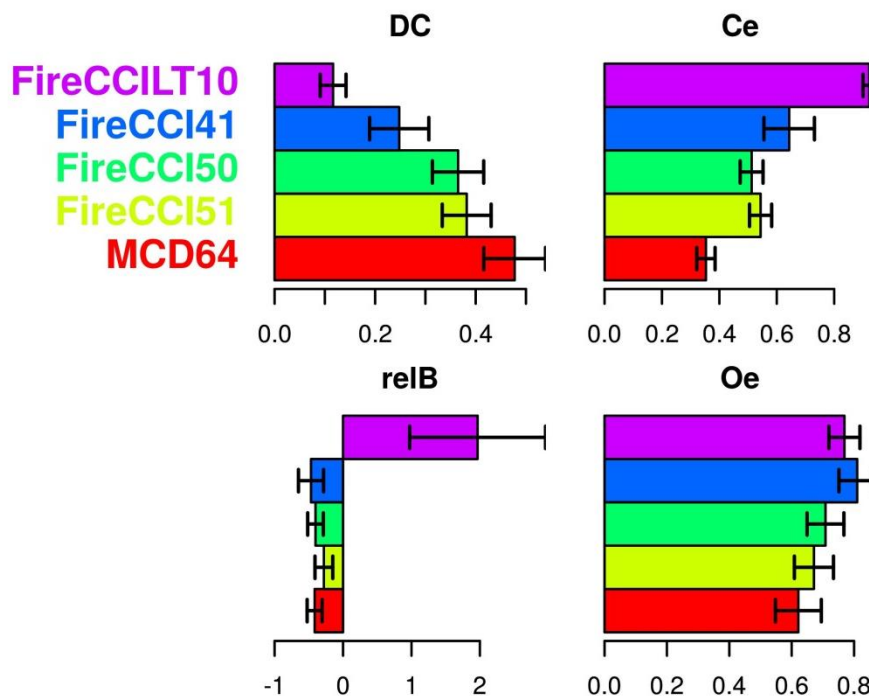


Figure 13: Estimated accuracy of each product. 95% confidence intervals are shown with the error segments.

Detailed accuracies on TSAs for Fire_cci products can be seen in Figures 14-17. BA at the reference data can be seen in Figure 18. 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 savannahs 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 Figures below), with the exception of FireCCILT10 with large overestimations in most TSAs. Similar trends can be observed with Ce and Oe, and for MCD64 (Annex 7).

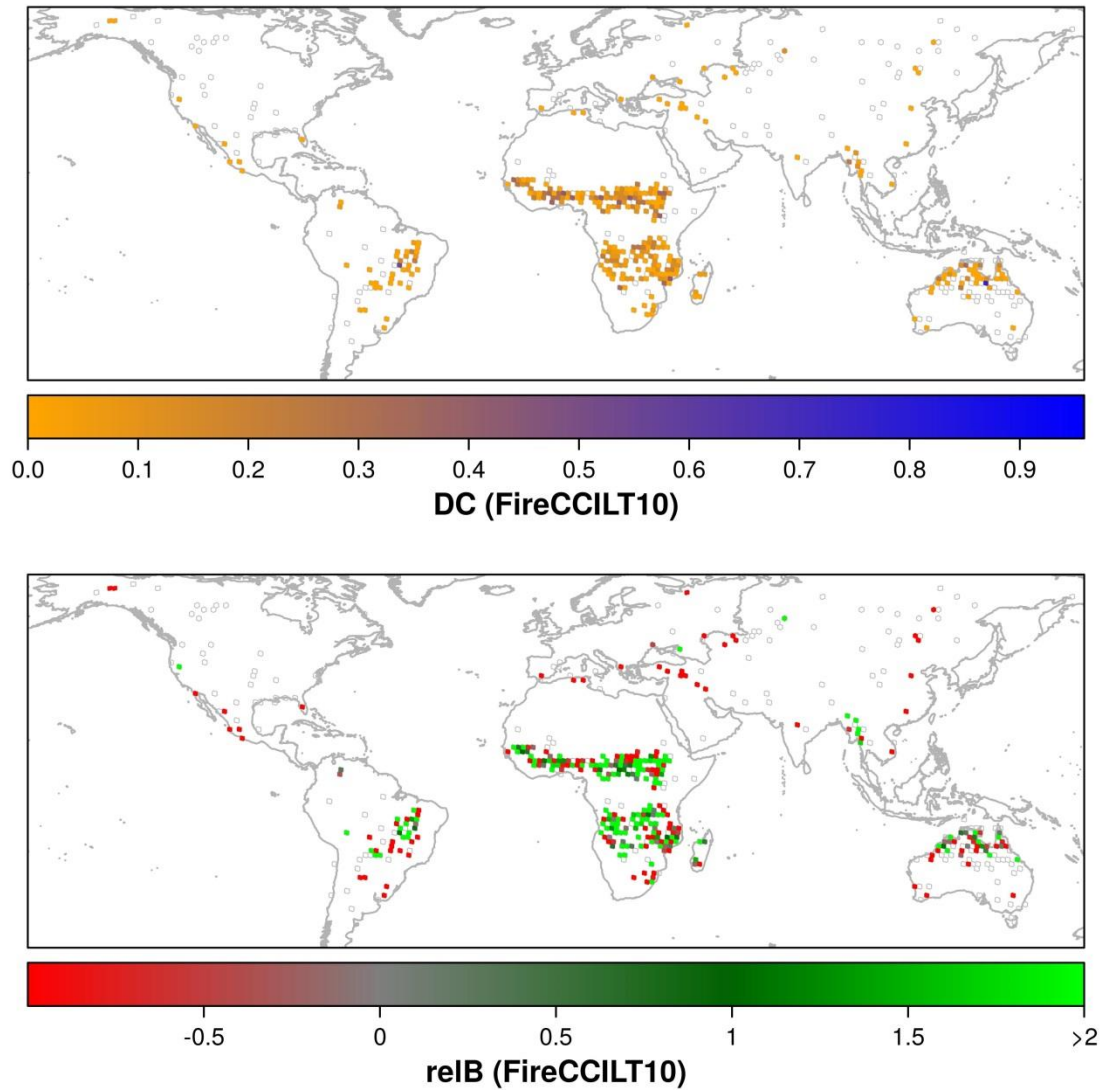


Figure 14: Dice Coefficient (DC) and relative bias (relB) for 2003-2014 FireCCI10 at TSAs. TSAs with reference data but without accuracy measures available are represented by empty polygons (white polygons with grey borders). DC is not available when there is no BA in the reference data nor in the product, and relB is not available when there is no BA in the reference data.

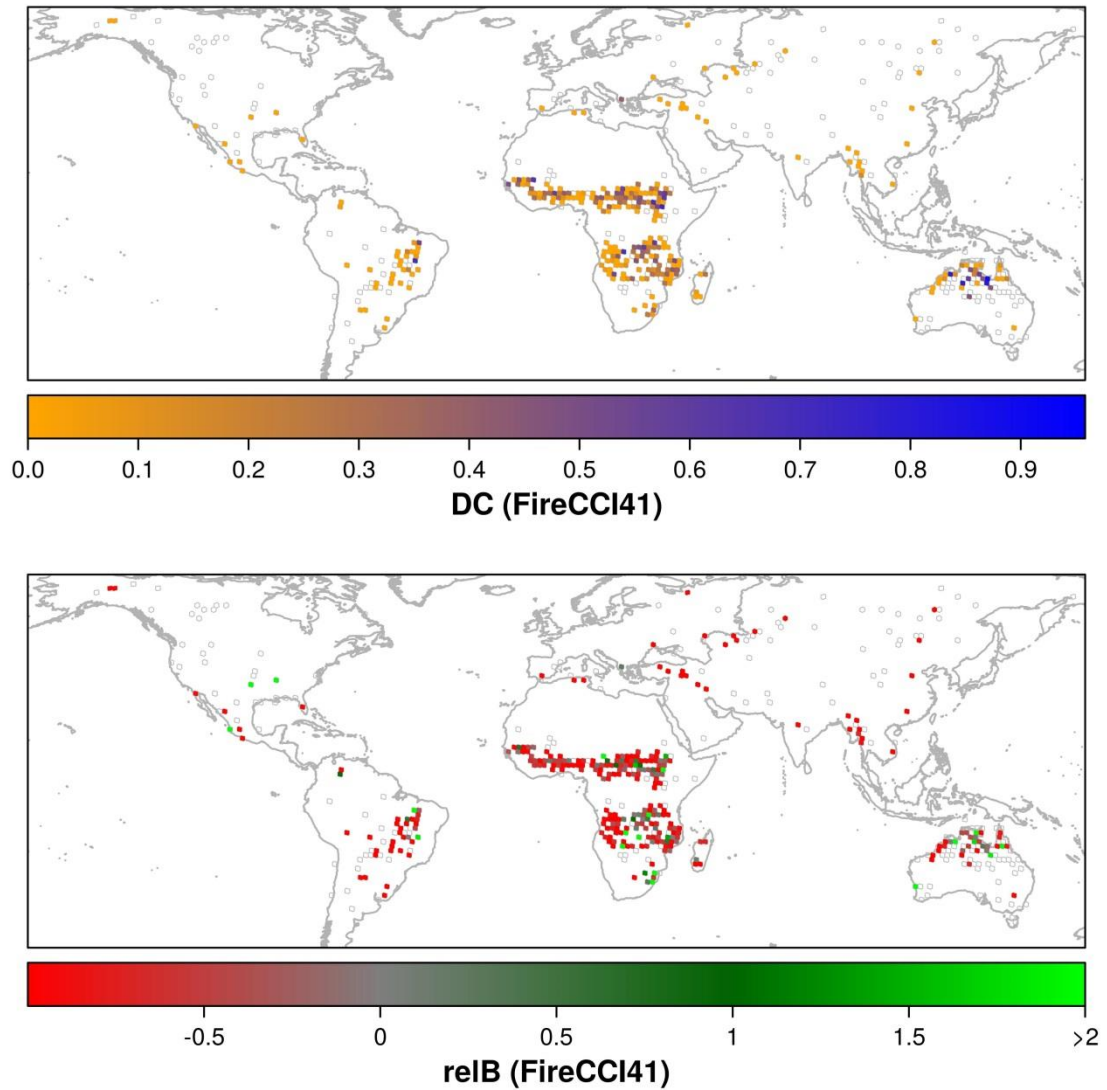


Figure 15: As in Figure 14 but for 2005-2011 FireCCI41.

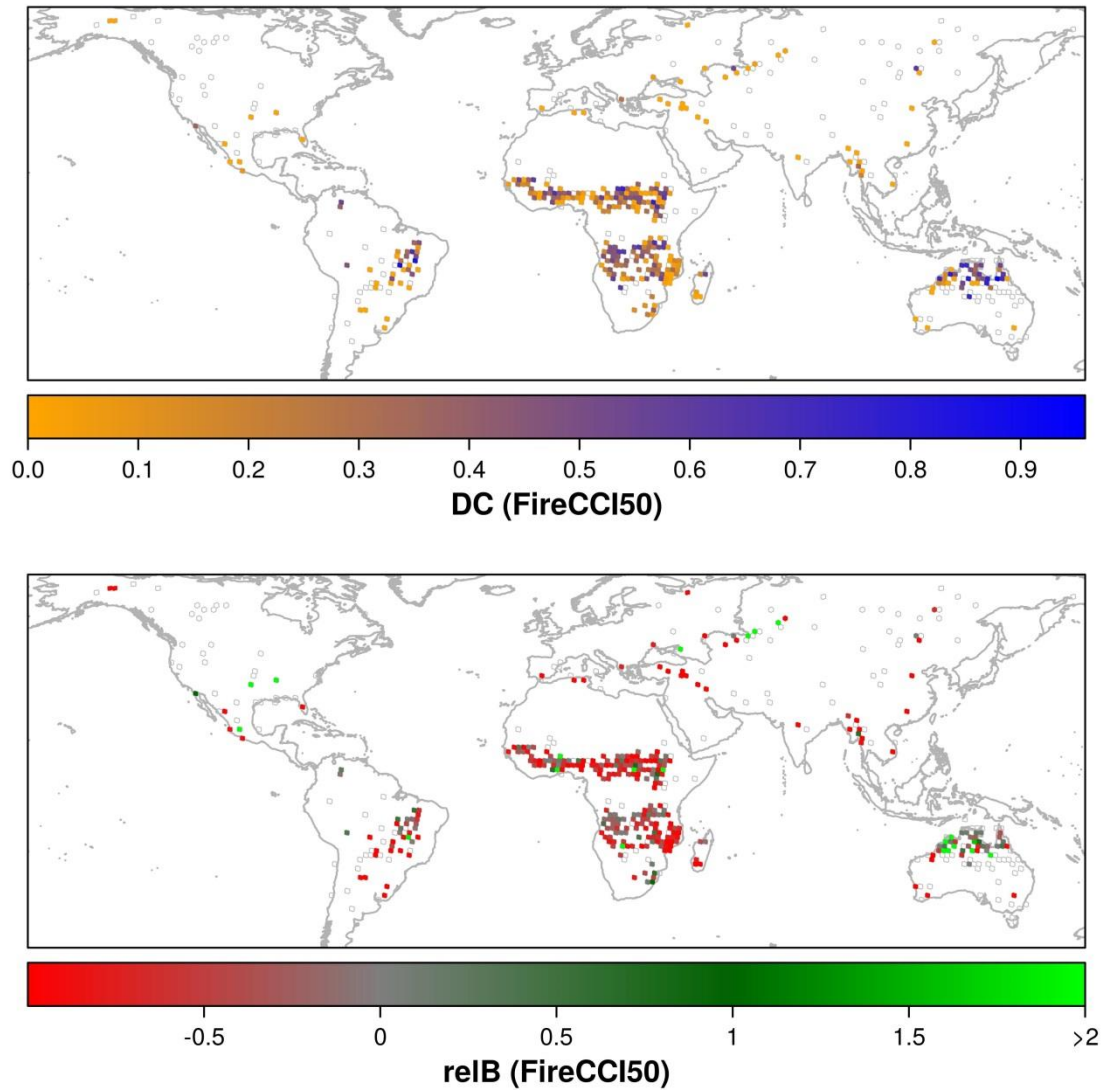


Figure 16: As in Figure 14 but for FireCCI50.

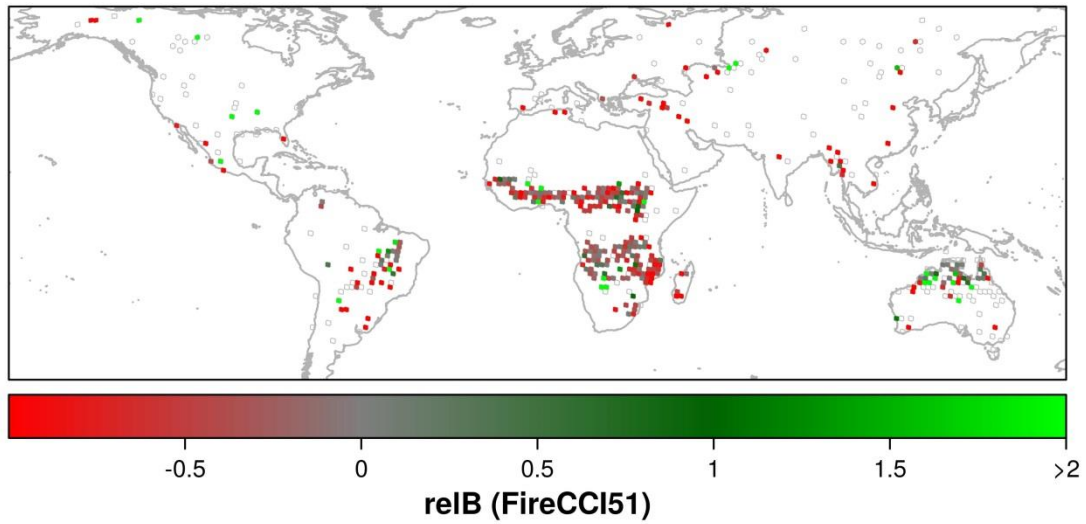
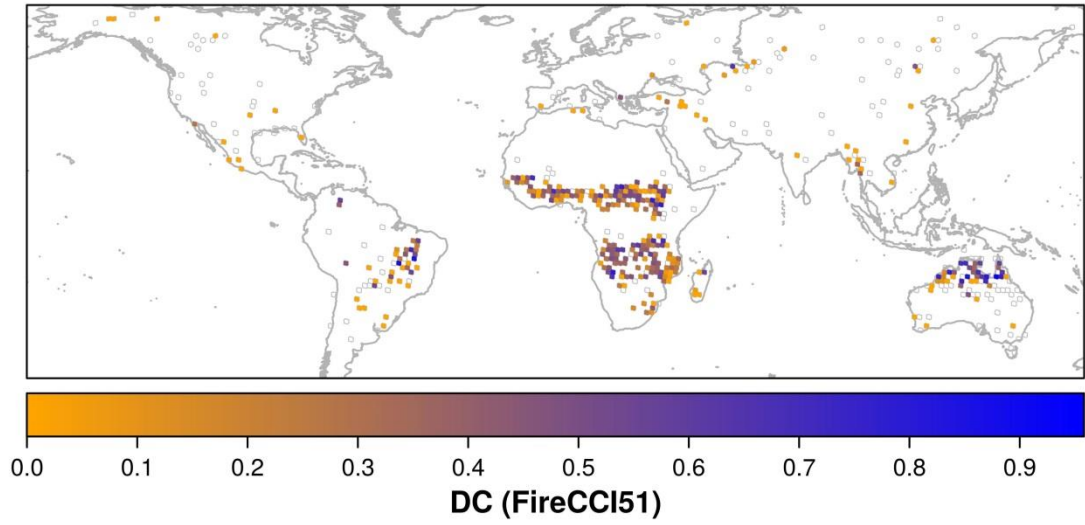


Figure 17: As in Figure 14 but for FireCCI51.

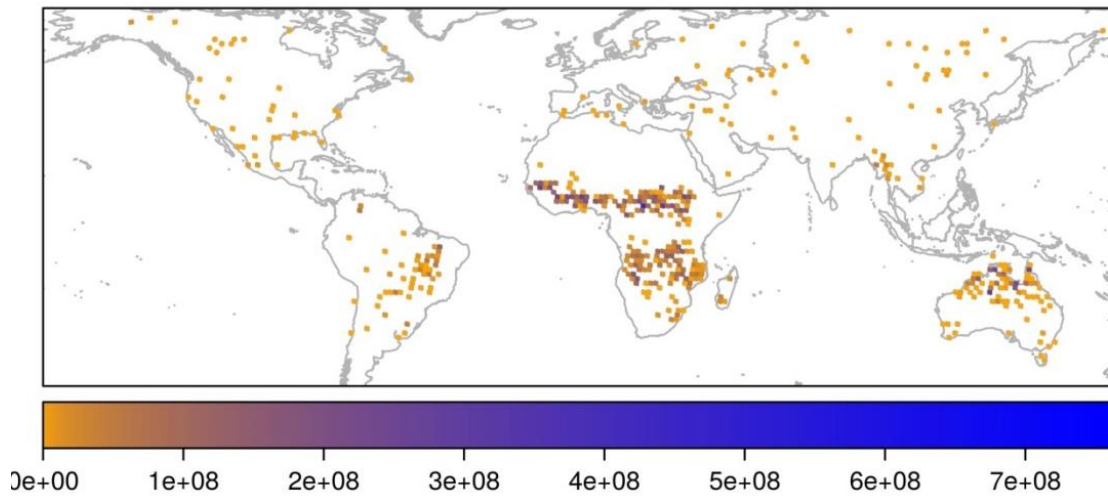


Figure 18: 2003-2014 BA (m²) in the reference data at TSAs.

Yearly accuracy estimates can be seen in Figure 19 and results of temporal monotonic trend tests in Table 4. A slight and steady increase in accuracy, although with a peak in the second year (2007), is observed for all products, with the exception of FireCCILT10. No significant temporal trends were detected.

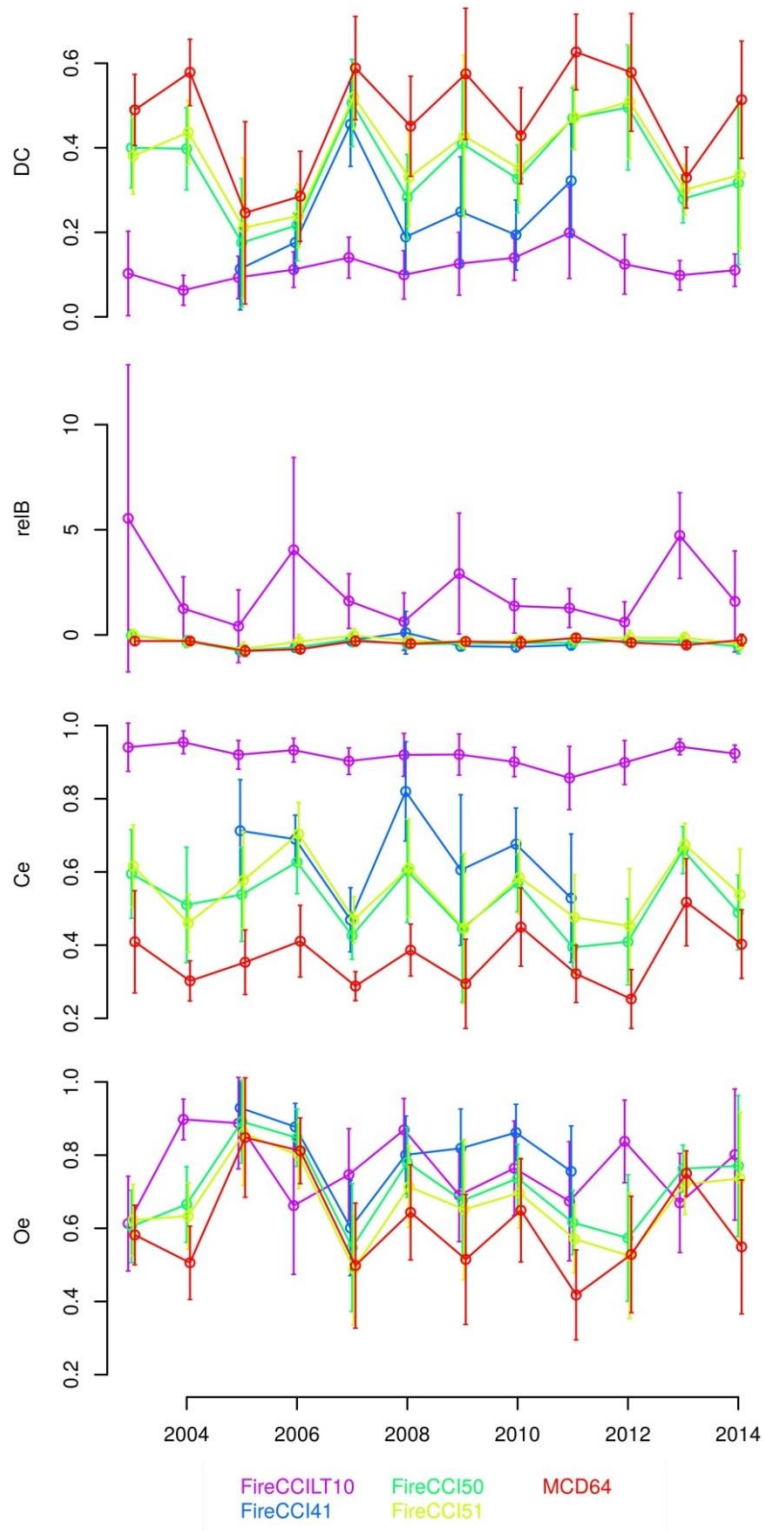


Figure 19: Yearly accuracy estimates. Vertical segments show the 95% confidence intervals.

Table 4: Temporal monotonic trends of accuracy (*b*; monotonic change of accuracy per year). None of them is significantly different from zero at $\alpha=0.05$ according Kendall's tau test.

	DC	relB	Ce	Oe
FireCCILT10	0.0035	-0.0425	-0.0032	-0.0057
FireCCI41	0.0252	0.0295	-0.0268	-0.0195
FireCCI50	0.0072	0.0032	-0.0088	-0.0001
FireCCI51	0.0083	0.0041	-0.0032	0.0025
MCD64	0.0048	0.0044	0.0027	-0.0044

4.2 SFD

FireCCISFD11 product is statistically the most accurate product at long sampling units (Figure 20), with the lowest commission error ratio (Oe), highest Dice Coefficient (DC) and lowest bias (relB). FireCCIS1A10 is the least accurate product with the smallest amount of data available (part of the northern hemisphere of sub-Saharan Africa) what lead to the largest uncertainties in accuracy estimates (reflected by the large standard errors and confidence intervals). Detailed results of error matrix entries (e_{ij}) and accuracy measures are presented as tables in Annex 8.

Accuracies at long sampling units were higher than those obtained at the scale of image pairs ("short units"), consistently in all products and particularly in FireCCISFD11, FireCCIS1A10, FireCCILT10, FireCCI51 and FireCCI50 (Figure 20). It is remarkable how FireCCISFD11, FireCCI51 and MCD64 have similar accuracy at long sampling units, even though the former two have slightly less accuracy than later at the short sampling units.

Bias remained unchanged between short and long units in all products. All products underestimate BA by around or more than 50% of what it is actually burned, except FireCCISFD11 and FireCCIS1A10 which have the lowest biases, the former underestimates 9% and the later overestimates 8%.

Detailed results of accuracy on long sampling units for Fire_cci products can be seen in Figures 21-Figure 25. BA at the reference data can be seen in Figure 26. Results with Ce and Oe and for MCD64 can be seen in Annex 9. Similarly as for the global sample 2003-2014, TSAs with the highest accuracies (i.e. highest DC) tend to be located where BA is high.

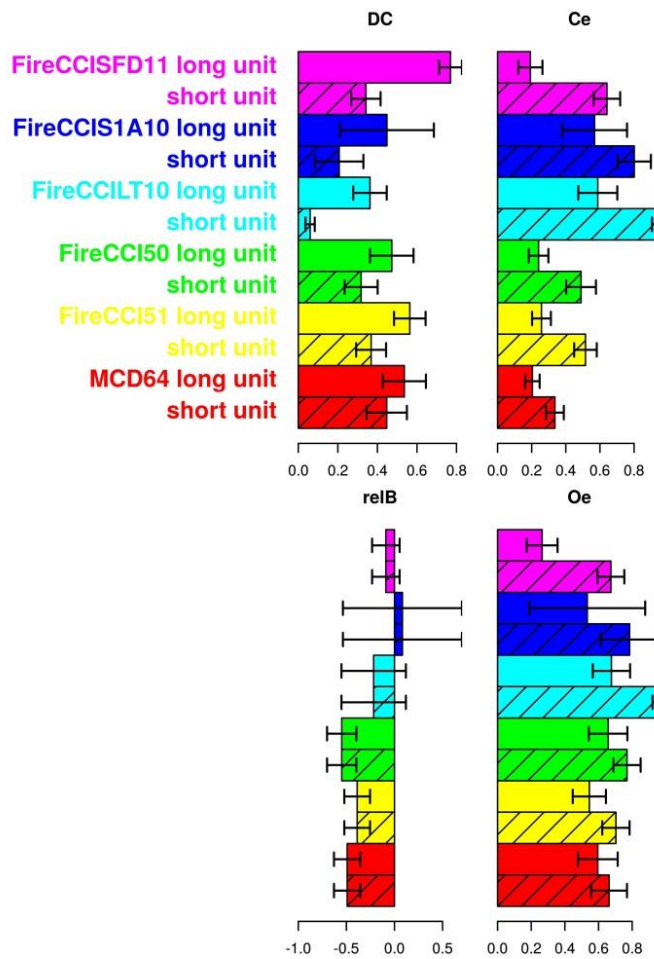


Figure 20: Estimated product accuracies at long sampling units (long su) and at the scale of image pairs (short su). 95% confidence intervals are shown with the error segments.

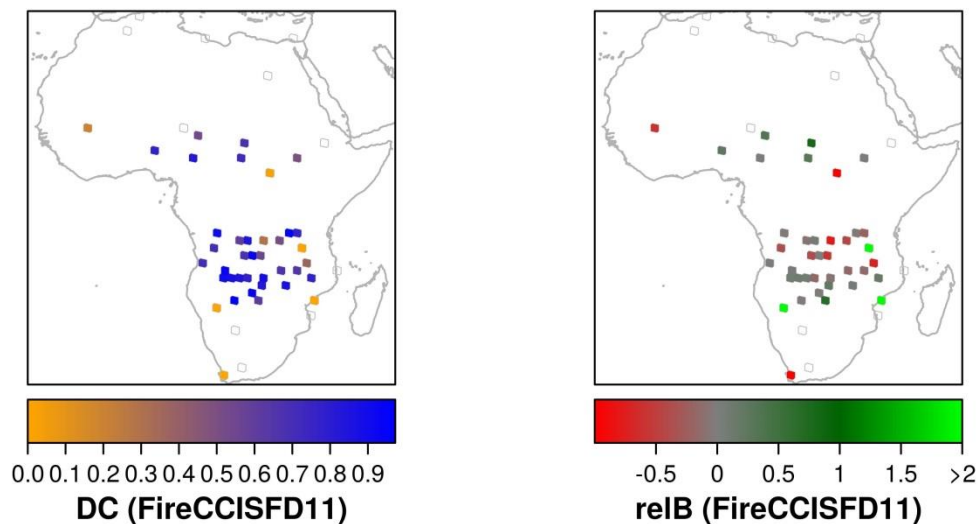


Figure 21: Dice of coefficient (DC) and relative bias (relB) for S2 FireCCISFD11 at TSAs for long sampling units over the sample of Africa 2016. Units without product data or without accuracy measures available are represented by empty polygons (white polygons with grey borders). DC is not available 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.



fire
cci

Fire_cci
Product Validation Report

Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
Issue	2.1	Date	22/12/2018
	Page		34

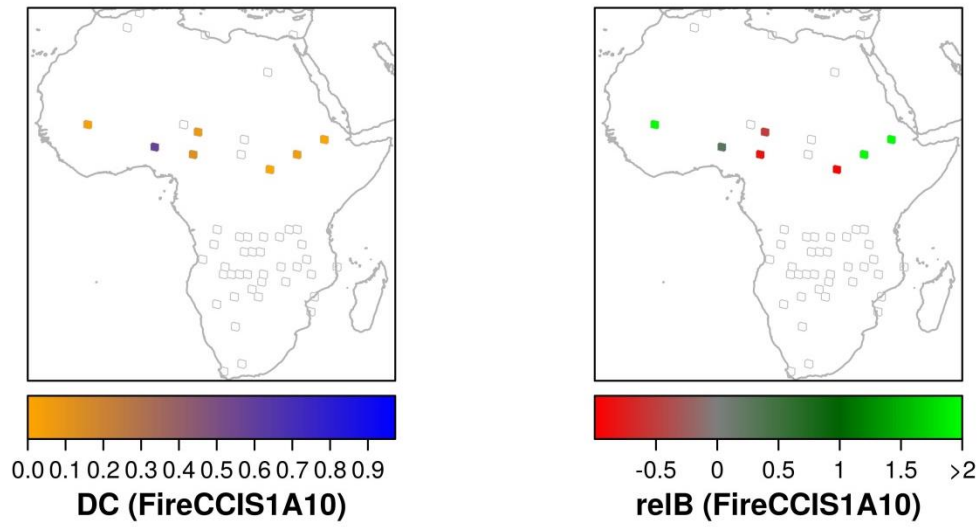


Figure 22: As in Figure 21 but for FireCCIS1A10.

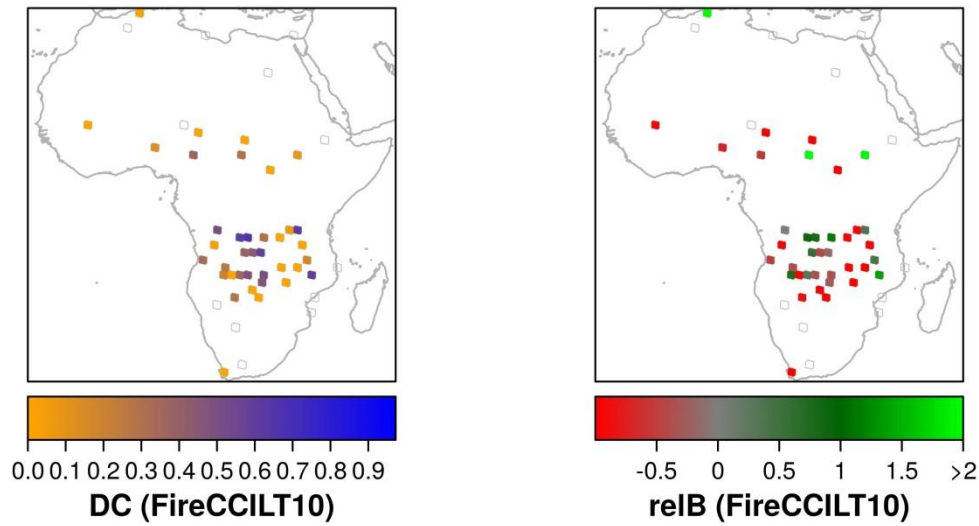


Figure 23: As in Figure 21 but for FireCCILT10.

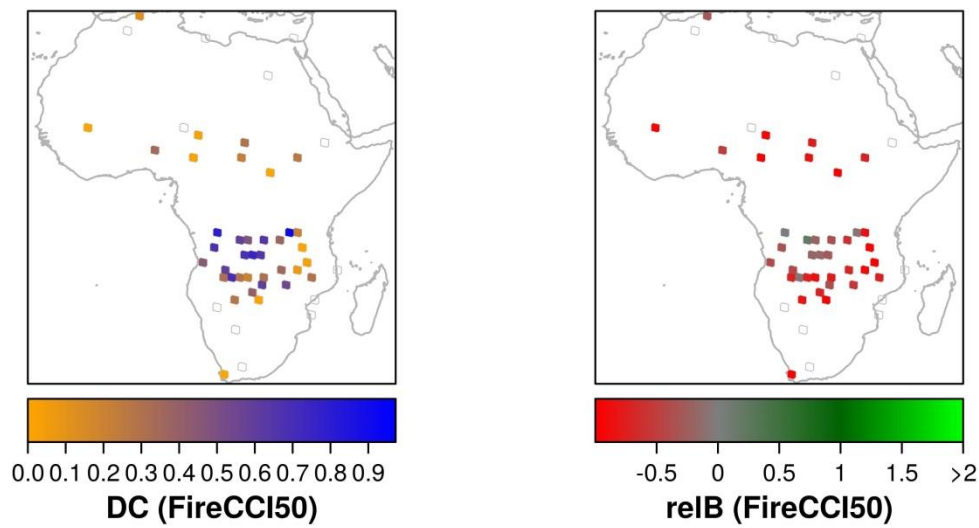


Figure 24: As in Figure 21 but for FireCCI50.

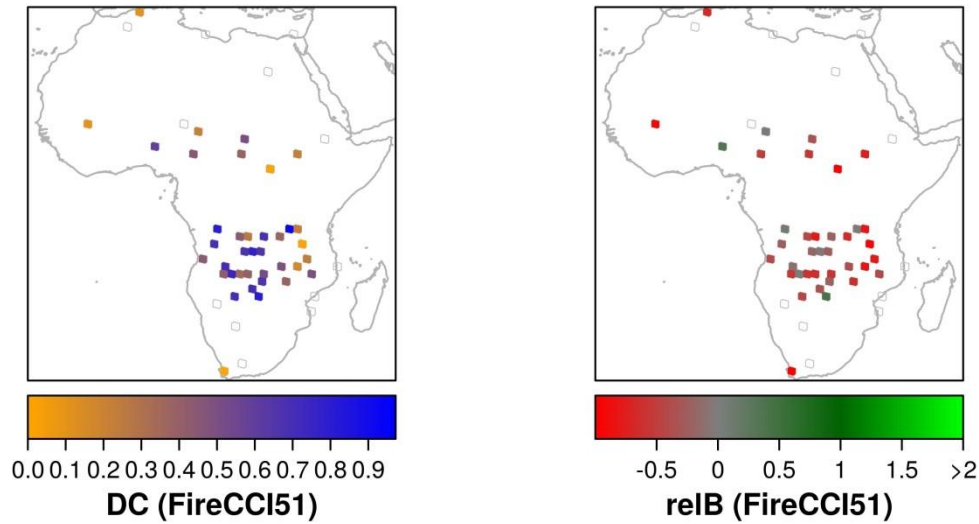


Figure 25: As in Figure 21 but for FireCCI51.

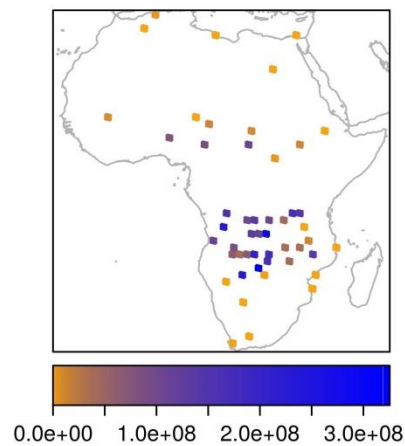


Figure 26: BA (m²) in the reference data at TSAs over the sample of Africa 2016.

It is remarkable how all global products have consistently low accuracy in some units while FireCCISFD11 has contrary higher accuracy, particularly due to a lower omission error ratio (Oe) across the whole Africa. This mainly occurs between the Sahara and the Equator. Figure 27 shows, for illustrative purposes, an example where a global product (FireCCI51, lower right panel) misses most of the BA in the reference data (pre-fire and post-fire times in the upper left and upper right panels), while FireCCISFD11 maps well most of the burns.

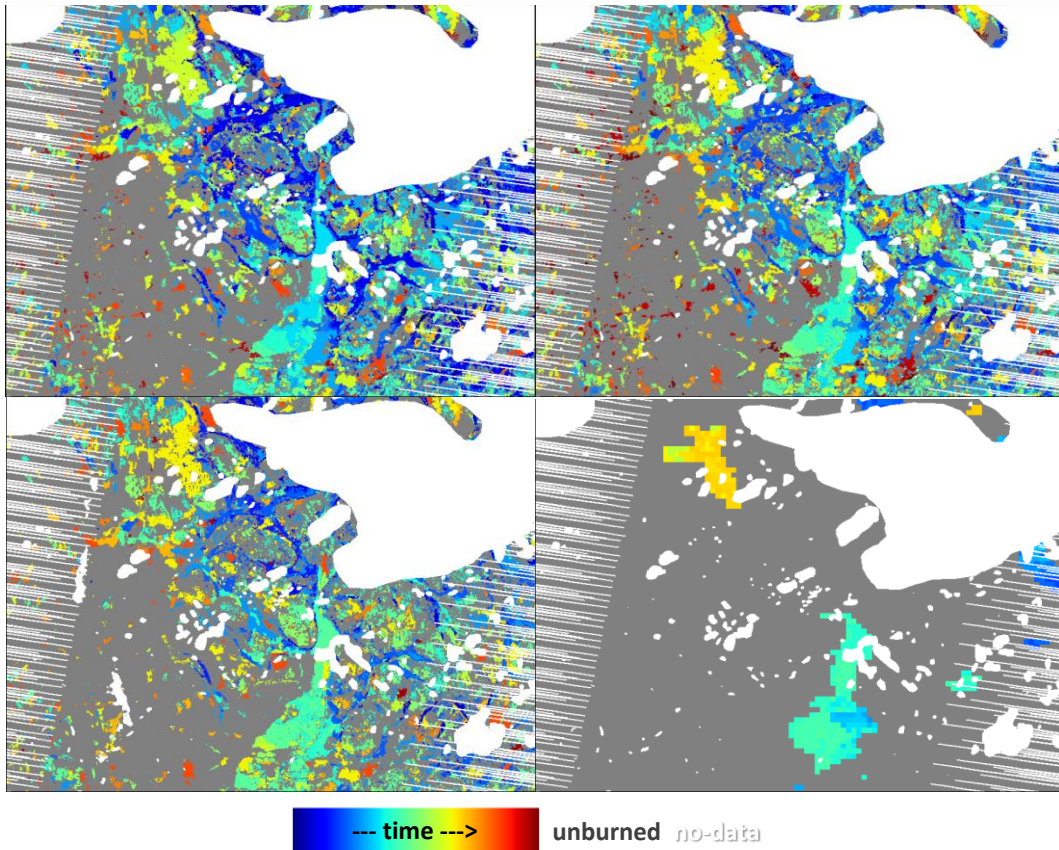



Figure 27: Areas burned at TSA path 170 and row 64 between 13 May and 20 October 2016, in the reference data (upper panels; identified by burning dates ranges, as in Figure 12), according to FireCCISFD11 (lower left) and according to FireCCI51 (lower right). Colours indicate the burn detection times. Grey represents unburned area and white no-data due to cloud coverage or SLC-OFF problems of ETM+.

5 Discussions and Conclusions

A new sampling design was developed to estimate the accuracy for long-time series of BA products. Validation results presented here are novel and the first ones belonging to the CEOS-LPVS validation stage 3. The part of the method that defines the stratification and sampling allocation has been published (Padilla et al., 2017), and was used as basis to develop new global and regional validation datasets for Fire_cci Phase 2, for global and SFD products. A total of 2252 multitemporal pairs of images were processed.

The very similar or even smaller standard errors of accuracy estimates compared with those from Phase 1 illustrate the efficiency 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 twelve years, while in Phase 1, global accuracy was available only for one year, 2008.

At global scale, FireCCI41, FireCCI50, FireCCI51, FireCCILT10 and MCD64 were validated and compared using reference BA data from Landsat TM generated at 1,200 sampling units distributed globally through twelve years (2003-2014). A stratified random sampling of spatio-temporal clusters on image pairs was used. Temporal stability of accuracy was assessed with available per-year accuracy estimates. Accuracy levels observed for FireCCI41 were similar to those observed in Phase 1 for v3.1 (Padilla and Chuvieco 2014).

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
					Page	37

FireCCI51 is currently the most accurate Fire_cci product; and less accurate than MCD64 particularly according the commission error ratio. This is at the short temporal interval, using consecutive image pairs of Landsat TM data for reference data. The accuracy results at long temporal reference data extents is different as reported below. FireCCILT10 is the least accurate product, probably due to the lower spatial resolution, different radiometric calibration, and lack of use of active fire information.

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 the 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 was expected and reflects that each algorithm uses the same input data throughout all the time period covered. The common temporal evolution of accuracy of products, with the exception of FireCCILT10, 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. This high covariance of accuracy between products was expected and already observed, in the spatial domain rather than in the temporal domain, in previous studies (Padilla et al. 2015). It is noteworthy that, even if precision of yearly accuracies is quite low, if one product has more accuracy than another, it consistently more accurate in the other years. Comparisons between product accuracies at different years must be done with caution, as samples are different (sampling units at different locations and time intervals; fires might have different characteristics).

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. This suggests that its impact on the temporal errors observed with current reference data accuracy is much lower than the variance of accuracy estimates, and that accuracies on areas with available data is similar to areas with not available data, as assumed to infer accuracies (Section 3.3).

In addition to the global validation dataset, the FireCCISFD11, FireCCIS1A10, FireCCILT10, FireCCI50, FireCCI51 and MCD64 products were validated using reference data at 50 long sampling units in Africa 2016. Each long sampling unit covers a time period of at least 100 days and is made by consecutive pairs of images (by the short sampling units). The differences of accuracy at long and short sampling units reflect how temporal reporting errors are less likely to be included as temporal extent of units increases. Such differences of accuracy were large in FireCCISFD11, FireCCIS1A10 and FireCCILT10. Those results were unexpected for FireCCILT10 but were expected for the former two products which can have larger temporal reporting errors due to the lower temporal resolution of their input data, compared to the around daily resolution input data common in global burned area products. The larger differences of accuracy at long and short units of FireCCI51 compared with MCD64 suggest that temporal reporting errors of the former are slightly higher than in the latter. The similar accuracies of those products at long units, the former slightly higher than the latter, at long sampling units reflects how two different approaches of using similar input data (MODIS reflectance and active fires) can lead to similar results at long temporal scales. Therefore, this implies that both products would be similarly suitable for a user that is not bothered about fire detection dates and that is mainly interested in



for example yearly BA maps. The higher accuracy of FireCCISFD11 at long sampling units than that of the global products, mainly in sites with low BA and due to less omission errors, reflects the benefits of a higher spatial resolution that allows for a better mapping of small and fragmented fires. The fragmentation of fires can occur in the spatial domain but also in the temporal domain, as illustrated by Figure 27. The succession of small fires over the days can appear as a smooth change on coarse spatial resolution pixels and not detected by the classification algorithm. Contrarily, those same small fires can appear as sudden changes at higher spatial resolution observations and be correctly detected by similar algorithms. The accuracy results on FireCCIS1A10 need more investigation as they were obtained with a small reference sample and uncertainties in accuracy estimates are large as shown in Figure 20.

6 References

- Bastarrika, A., Chuvieco, E., & Martin, M.P. (2011). 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
- Boschetti, L., Roy, D., & Justice, C. (2009). International Global Burned Area Satellite Product Validation Protocol. Part I – production and standardization of validation reference data. In C. CalVal (Ed.) (pp. 1-11). USA: Committee on Earth Observation Satellites
- Boschetti, L., Roy, D.P., Justice, C.O., 2010. International global burned area satellite product validation protocol part i — production and standardization of validation reference data. Tech. Rep. CEOS Working Group on Calibration and Validation. <https://lpvs.gsfc.nasa.gov/PDF/BurnedAreaValidationProtocol.pdf>
- Boschetti, L., Roy, D.P., Justice, C.O., & Giglio, L. (2010). Global assessment of the temporal reporting accuracy and precision of the MODIS burned area product. *International Journal of Wildland Fire*, 19, 705-709
- Boschetti, L., Stehman, S., V., & Roy, D.P. (2016). 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
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32
- Chuvieco, E., Opazo, S., Sione, W., Valle, H. d., Anaya, J., Bella, C. D., 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. (2008), Global burned-land estimation in Latin America using MODIS composit data. *Ecological Applications*, 18: 64-79. doi:10.1890/06-2148.1
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons
- Cohen, W.B., Yang, Z., & Kennedy, R.E. (2010). 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
- Congalton, R.G., & Green, K. (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Applications*. Boca Raton: Lewis Publishers
- Conover, W.J. (1999). *Practical Nonparametric Statistics*. John Wiley & Sons
- Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26, 297-302
- European Space Agency (2007). Earth Observation Satellites: A Roadmap For Calibration And Validation. ScienceDaily. ScienceDaily, 29 October 2007. <www.sciencedaily.com/releases/2007/10/071022120154.htm>.
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. Canada: John Wiley & Sons



- Forbes, A.D. (1995). Classification-algorithm evaluation: five performance measures based on confusion matrices. *Journal of Clinical Monitoring*, 11, 189-206
- Gallego, F.J. (2005). Stratified sampling of satellite images with a systematic grid of points *Journal of Photogrammetry & Remote Sensing*, 59, 369-376
- GCOS (2016). The Global Observing System for Climate: Implementation Needs. In: World Meteorological Organization
- Giglio, L., Boschetti, L., Roy, D.P., Humber, M.L., & Justice, C.O. (2018). The Collection 6 MODIS burned area mapping algorithm and product. *Remote Sensing of Environment*, 217, 72-85
- Giglio, L., Loboda, T., Roy, D.P., Quayle, B., & Justice, C.O. (2009). An active-fire based burned area mapping algorithm for the MODIS sensor. *Remote Sensing of Environment*, 113, 408-420
- Giglio, L., Randerson, J., T., van der Werf, G.R., Kasibhatla, P., Collatz, G.J., Morton, D.C., & Defries, R. (2010). Assessing variability and long-term trends in burned area by merging multiple satellite fire products. *Biogeosciences Discuss*, 7, 1171
- Goodwin, N.R., & Collet, L.J. (2014). Development of an automated method for mapping fire history captured in Landsat TM and ETM+ time series across Queensland, Australia. *Remote Sensing of Environment*, 148, 206-221
- Hand, D.J. (1981). *Discrimination and Classification*. New York: John Wiley and Sons
- Hansen, H.M., Hurwitz, W.N., & Gurney, M. (1946). Problems and Methods of the Sample Survey of Business. *Journal of the American Statistical Association*, 41, 173-189
- Hantson, S., Padilla, M., Corti, D., & Chuvieco, E. (2013). Strengths and weaknesses of MODIS hotspots to characterize global fire occurrence. *Remote Sensing of Environment*, 131, 152-159
- Heil, A., Kaiser, J.W., Bistinas, I., & van der Werf, G. (2017). ESA CCI ECV Fire Disturbance: D.4.1.2. Product Intercomparison Report, version 1.0. In F.c. project (Ed.)
- Heil, A., Yue, C., Mouillot, F., & Kaiser, J.W. (2016). ESA CCI ECV Fire Disturbance: D1.1 User requirements document, version 5.1. In F.c. project (Ed.)
- Hellden, U. (1980). A test of Landsat-2 imagery and digital data for thematic mapping, illustrated by an environmental study in northern Kenya. In. Sweden: Lund University Natural Geography Institute
- Kennedy, R.E., Yang, Z., & Cohen, W.B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr - Temporal segmentation algorithms. *Remote Sensing of Environment*, 114, 2897-2910
- Latifovic, R., & Olthof, I. (2004). Accuracy assessment using sub-pixel fractional error matrices of global land cover products derived from satellite data. *Remote Sensing of Environment*, 90, 153-165
- Liu, C., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107, 606-616
- Morisette, J.T., Baret, F., & Liang, S. (2006). Special issue on global land product validation. *IEEE Trans. Geosci. Remote Sens.* 44, 1695-1697.
- Mouillot, F., Schultz, M.G., Yue, C., Cadule, P., Tansey, K., Ciais, P., & Chuvieco, E. (2014). Ten years of global burned area products from spaceborne remote sensing - A review: Analysis of user needs and recommendations for future developments. *International Journal of Applied Earth Observation and Geoinformation*, 26, 64-79
- 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., & Kassem, K.R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, 51, 933-938
- Padilla, M., & Chuvieco, E. (2014). ESA CCI ECV Fire Disturbance - Product Validation Report II – Results from the Global Sample. In: ESA Fire-CCI project
- Padilla, M., Chuvieco, E., Hantson, S., Theis, R., & Snadow, C. (2014a). ESA CCI ECV Fire Disturbance - Product Validation Plan. In: ESA Fire-CCI project
- Padilla, M., Olofsson, P., Stephen V., S., Tansey, K., & Chuvieco, E. (2017). Stratification and sample allocation for reference burned area data. *Remote Sensing of Environment*, 203, 240-255
- Padilla, M., Stehman, S.V., & Chuvieco, E. (2014b). Validation of the 2008 MODIS-MCD45 global burned area product using stratified random sampling. *Remote Sensing of Environment*, 144, 187-196
- Padilla, M., Stehman, S.V., Litago, J., & Chuvieco, E. (2014c). Assessing the temporal stability of the accuracy of a time series of burned area products. *Remote Sensing*, 6, 2050-2068
- Padilla, M., Stehman, S.V., Ramo, R., Corti, D., Hantson, S., Oliva, P., Alonso, I., Bradley, A., Tansey, K., Mota, B., Pereira, J.M., & Chuvieco, E. (2015). Comparing the Accuracies of Remote Sensing Global Burned Area Products using Stratified Random Sampling and Estimation. *Remote Sensing of Environment*, 160, 114-121
- Pedregosa, F.V., G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830
- Plummer, S., Arino, O., Ranera, F., Tansey, K., Chen, J., Dedieu, G., Eva, H., Piccolini, I., Leigh, R., Borstlap, G., Beusen, B., Fierens, F., Heyns, W., Benedetti, R., Lacaze, R., Garrigues, S., Quaipe, T., De Kauwe, M., Quegan, S., Raupach, M., Briggs, P., Poulter, B., Bondeau, A., Rayner, P., Schultz, M., & McCallum, I. (2007). An update on the GlobCarbon initiative: multi-sensor estimation of global biophysical products for global terrestrial carbon studies. In, *Envisat Symposium 2007*. Montreux, Switzerland
- Ramo, R., & Chuvieco, E. (2017). Developing a Random Forest Algorithm for MODIS Global Burned Area Classification. *Remote Sensing*, 9, 1193
- Randerson, J.T., Chen, Y., van der Werf, G.R., Rogers, B.M., & Morton, D.C. (2012). Global burned area and biomass burning emissions from small fires *Journal of Geophysical Research*, 117
- Roy, D.P., & Boschetti, L. (2009). Southern Africa validation of the MODIS, L3JRC, and GlobCarbon burned-area products. *IEEE Transactions on Geoscience and Remote Sensing*, 47, 1032-1044
- Roy, D.P., Boschetti, L., Justice, C.O., & Ju, J. (2008). The collection 5 MODIS burned area product - Global evaluation by comparison with the MODIS active fire product. *Remote Sensing of Environment*, 112, 3690-3707
- Stehman, S.V. (1997). Estimating Standard Errors of Accuracy Assessment Statistics under Cluster Sampling. *Remote Sensing of Environment*, 60, 258-269
- Tansey, K., Grégoire, J.-M., Defourny, P., Leigh, R., Pekel, J.-F., Bogaert, E., & Bartholome, E. (2008). A new, global, multi-annual (2000-2007) burnt area product at 1 km resolution. *Geophysical Research Letters*, 35, L01401, doi:10.1029/2007GL03156

	Fire_cci Product Validation Report	Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
		Issue	2.1	Date	22/12/2018
				Page	41

Wessels, K., van den Bergh, F., Roy, D., Salmon, B., Steenkamp, K., MacAlister, B., Swanepoel, D., & Jewitt, D. (2016). Rapid Land Cover Map Updates Using Change Detection and Robust Random Forest Classifiers. *Remote Sensing*, 8, 888

Annex 1 Acronyms and abbreviations

ABAMS	Automatic Burned Area Mapping Software
BA	Burned Area
CalVal	Calibration and Validation
CBERS	China-Brazil Earth Resources Satellite
Ce	Commission error ration
CEOS	Committee on Earth Observation Satellites
DC	Dice Coefficient
dNBR	Difference Normalized Burn Ratio
ESA	European Space Agency
ESRI	Environmental Systems Research Institute
ECV	Essential Climate Variables
ETM+	Enhanced Thematic Mapper +
FireCCI41	MERIS Fire_cci v4.1
FireCCI50	MODIS Fire_cci v5.0
FireCCI51	MODIS Fire_cci v5.1
FireCCILT10	AVHRR-LTDR Fire_cci v1.0
FireCCIS1A10	Sentinel-1 Fire_cci for Africa v1.0
FireCCISFD11	Sentinel-2 Fire_cci v1.1
GCOS	Global Climate Observing System
GFED3	Global Fire Emission Database v.3
L3JRC	Global Multi-year (2000-2007) Validated Burnt Area Product
LPV	Land Product Validation Subgroup of CEOS
OLI	Operational Land Imager
MCD45	MODIS Collection 5 Burned Area product using the Roy et al. (2008) algorithm
MCD64	MODIS Collection 5 Burned Area product using the Giglio et al. (2009) algorithm
MERIS	Medium Resolution Imaging Spectrometer
MODIS	Moderate Resolution Imaging Spectroradiometer
NBR	Normalized Burn Ratio
NIR	Near InfraRed
Oe	Omission error ration
OLI	Operational Land Imager
PVR	Product Validation Report
relB	Relative bias
RGB	Red-Green-Blue composite
S-1	Sentinel-1
S-2	Sentinel-2
SLC	Scan Line Corrector
SFD	Small Fire Dataset
SWIR	Short Wave InfraRed
TM	Thematic Mapper
TSA	Thiessen Scene Area
UTM	Universal Transverse Mercator
WGS84	World Geodetic System 1984
WRS(-2)	Worldwide Reference System (version 2)
XML	eXtensible Markup Language



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 classified 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 plugin

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.

	Fire_cci Product Validation Report		Ref.:	Fire_cci_D4.1.1_PVR_v2.1		
			Issue	2.1	Date	22/12/2018
					Page	45

Annex 4 Example of a XML metadata file

```

<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-cci.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 $n_{yb} < 4$
 - o $n_{yb} = 4$ (n_{yb} is forced to be four)
 - o $BA_{yb} = 0$ (BA_{yb} is forced to be zero)
- Recalculation of $n_y = n_y - n$ added in the previous step
- 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 Population estimates of error matrix entries (e_{ij}) and accuracy measures for the global sample 2003-2014

Table 5: Estimated error matrices and reference burned area (m²) for each product. Standard errors of the estimates are shown in parentheses.

	e_{11}	e_{12}	e_{21}	e_{22}	BA_{ref}
FireCCILT10	3.42e+13 (3e+12)	4.05e+14 (7e+13)	1.14e+14 (1e+13)	5.45e+16 (6e+15)	1.48e+14 (1e+13)
FireCCI41	1.88e+13 (3e+12)	3.4e+13 (7e+12)	8.04e+13 (1e+13)	3.34e+16 (5e+15)	9.92e+13 (1e+13)
FireCCI50	4.35e+13 (4e+12)	4.57e+13 (4e+12)	1.06e+14 (1e+13)	5.49e+16 (6e+15)	1.49e+14 (1e+13)
FireCCI51	4.9e+13 (5e+12)	5.84e+13 (6e+12)	1e+14 (1e+13)	5.49e+16 (6e+15)	1.49e+14 (1e+13)
MCD64	5.85e+13 (5e+12)	3.19e+13 (3e+12)	9.6e+13 (1e+13)	4.41e+16 (6e+15)	1.55e+14 (2e+13)

Table 6: Estimated accuracy of each product. Standard errors of the estimates are shown in parentheses.

	DC	relB	Ce	Oe
FireCCILT10	0.116 (0.013)	1.966 (0.506)	0.922 (0.011)	0.769 (0.025)
FireCCI41	0.248 (0.030)	-0.468 (0.094)	0.643 (0.045)	0.810 (0.030)
FireCCI50	0.365 (0.026)	-0.402 (0.058)	0.512 (0.020)	0.708 (0.030)
FireCCI51	0.382 (0.025)	-0.280 (0.066)	0.544 (0.020)	0.671 (0.032)
MCD64	0.478 (0.031)	-0.415 (0.056)	0.353 (0.016)	0.622 (0.038)

Annex 7 Accuracy observations at TSAs for the global sample 2003-2014

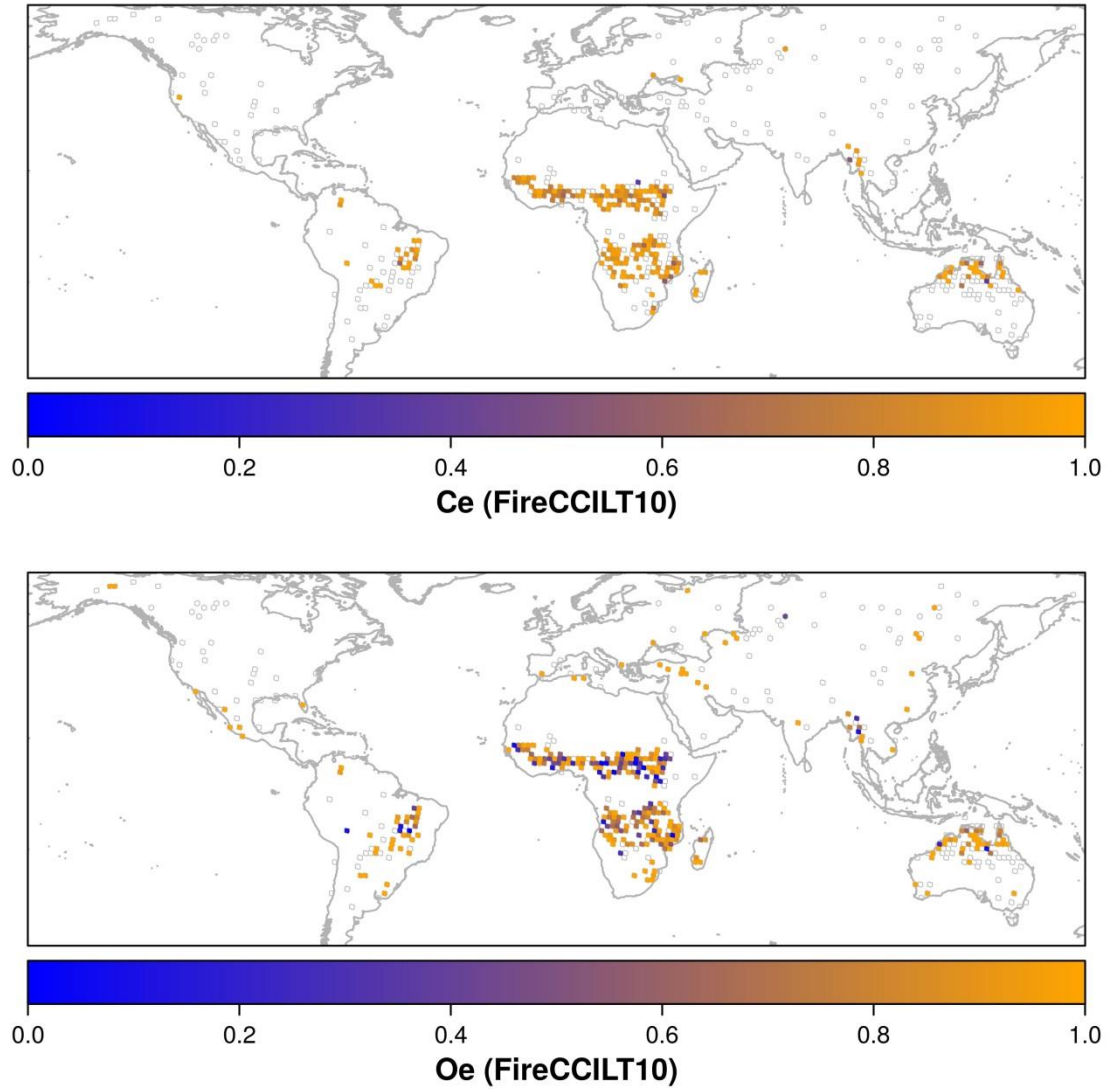


Figure 28: Ce and Oe at TSAs for FireCCILT10. TSAs with reference data but without accuracy measure available are represented by empty polygons (white polygons with grey borders).

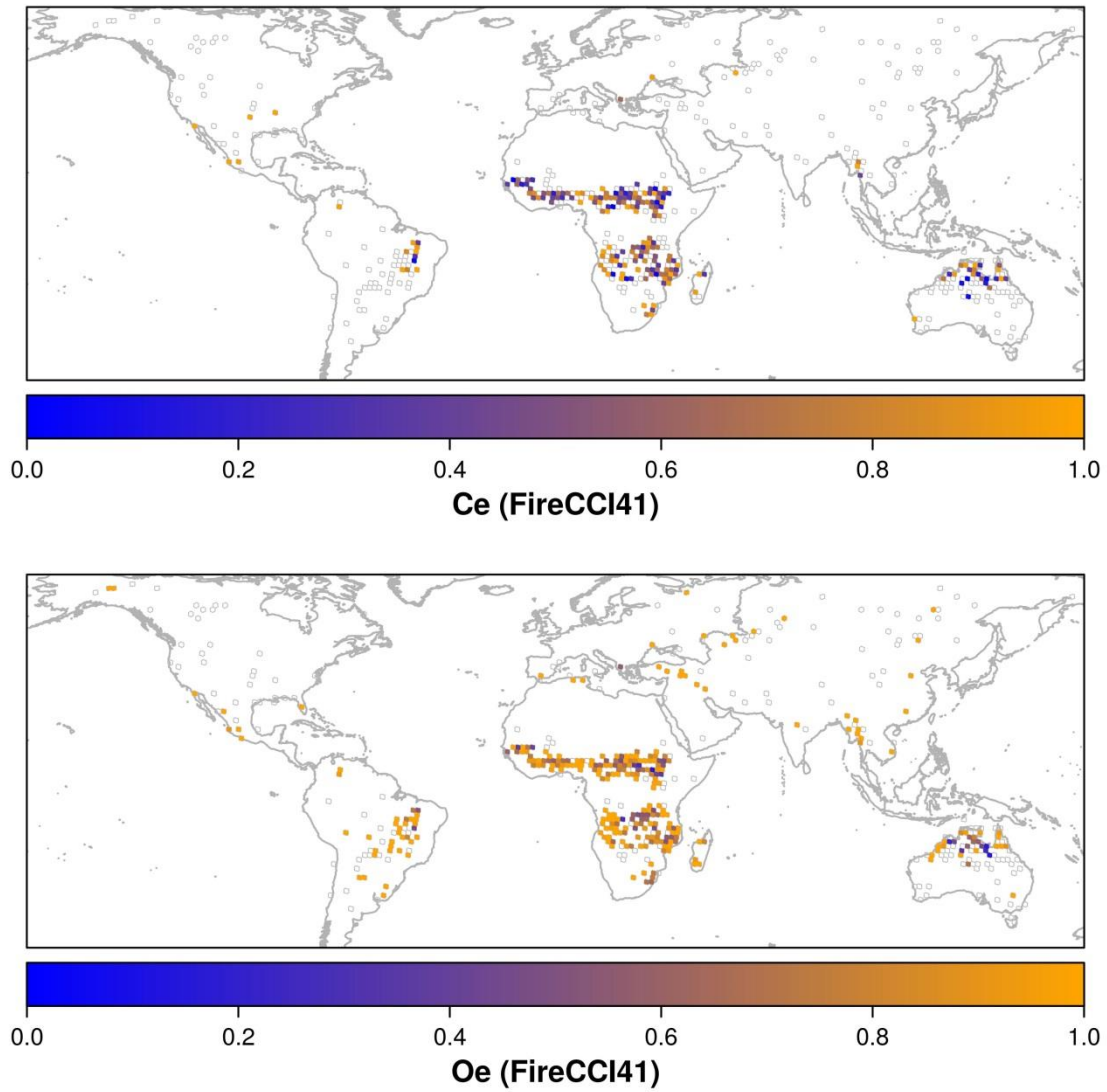


Figure 29: As in Figure 28 but for FireCCI41.

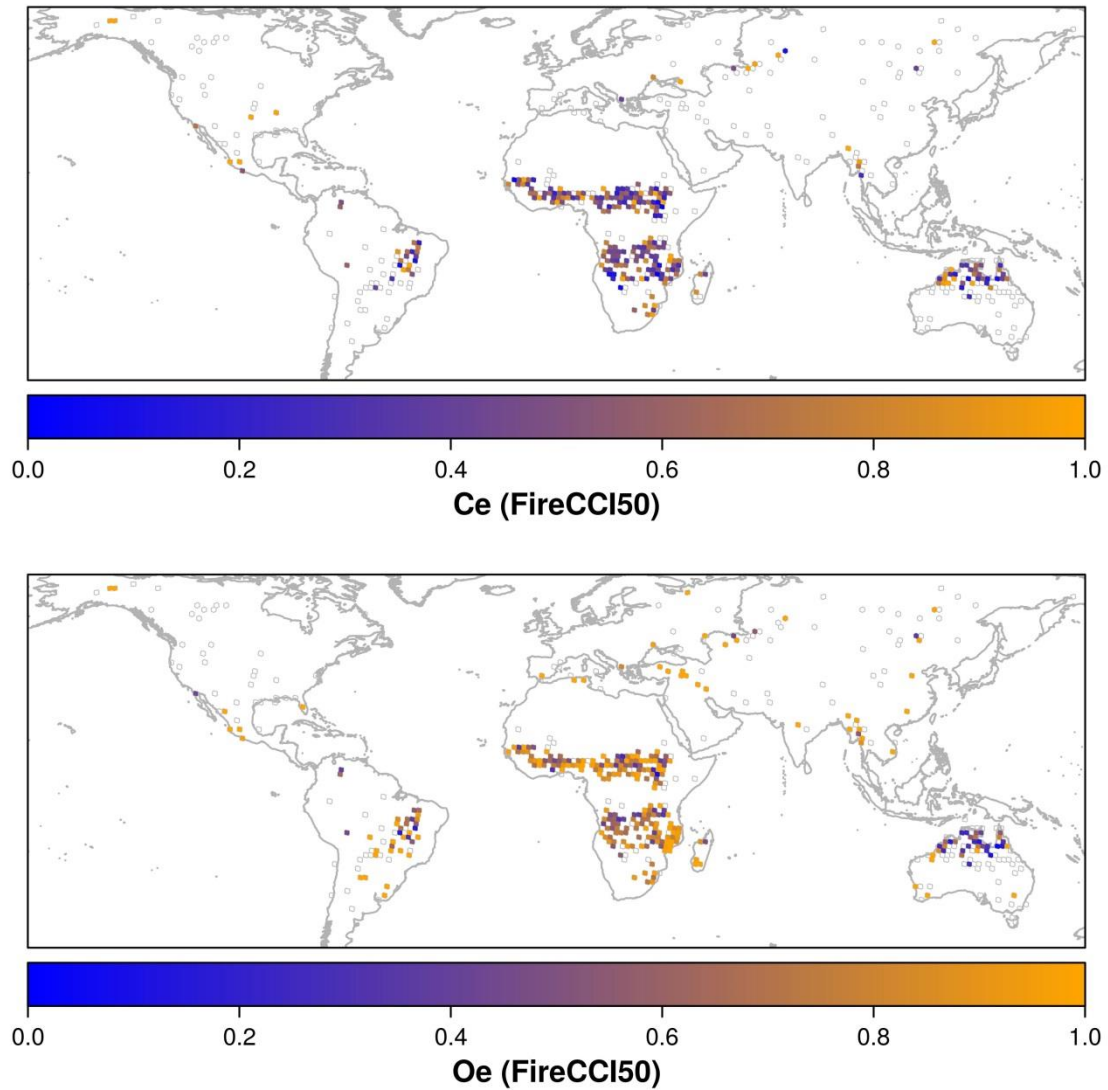


Figure 30: As in Figure 28 but for FireCCI50.

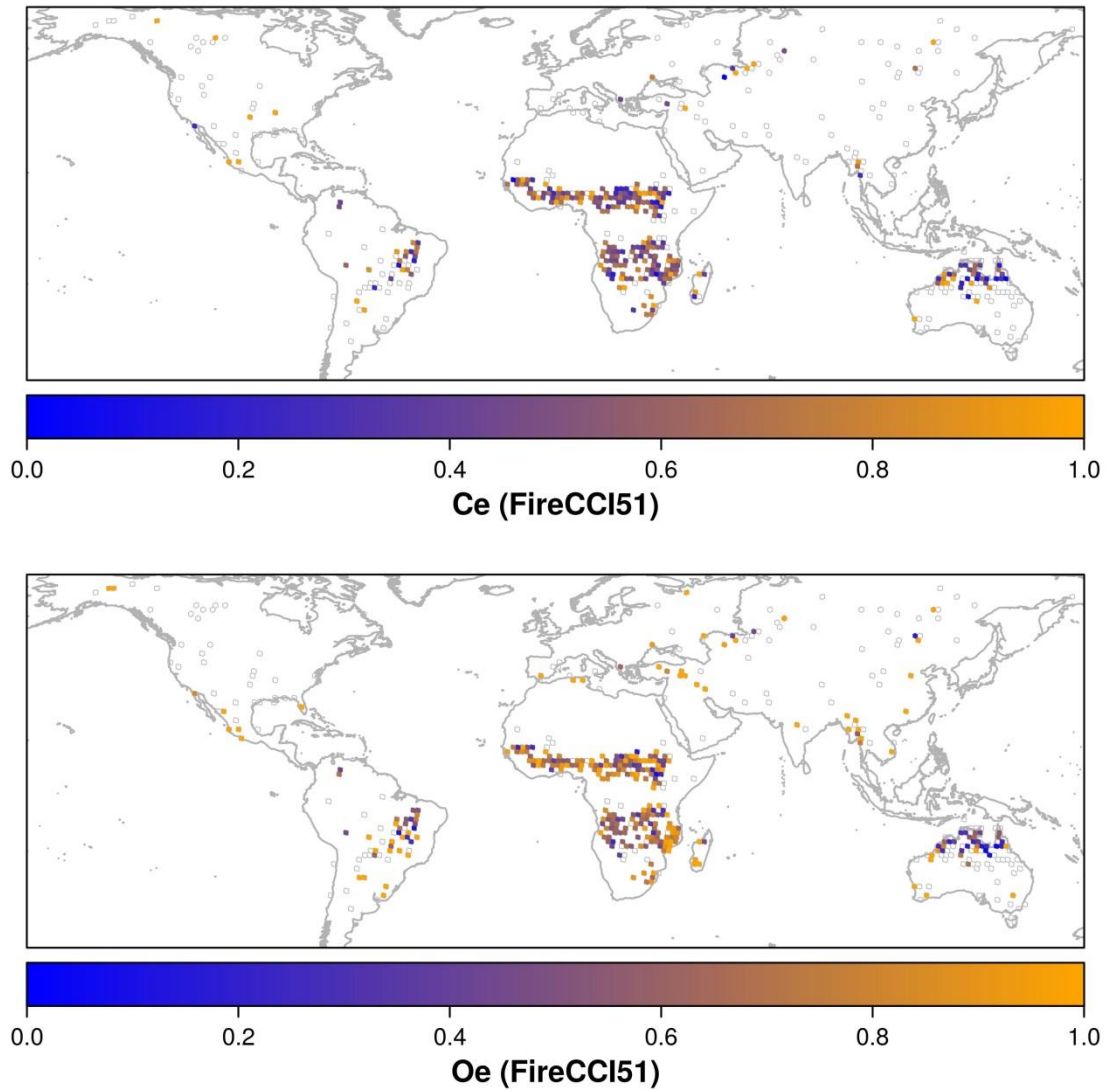


Figure 31: As in Figure 28 but for FireCCI51.

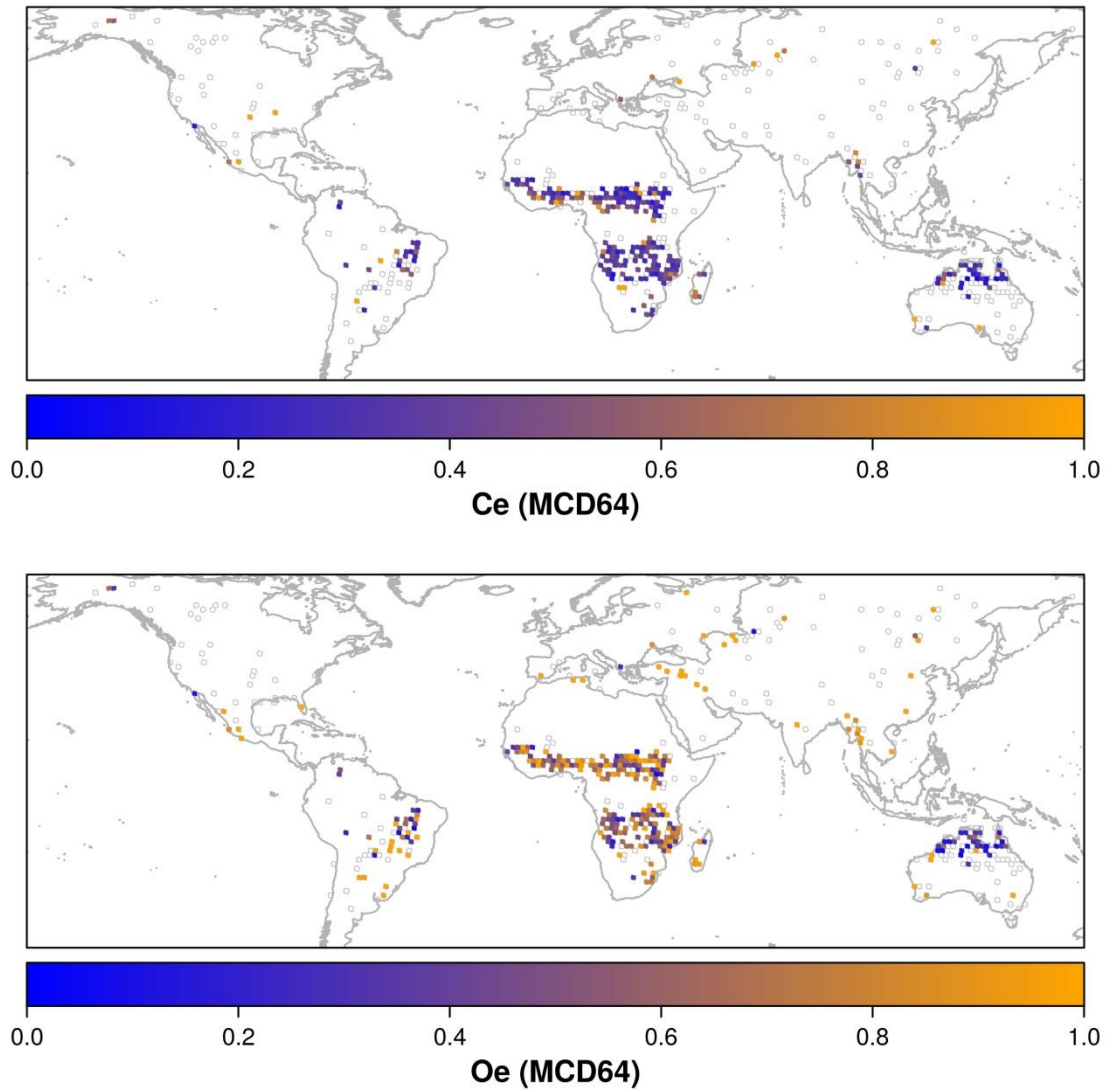


Figure 32: As in Figure 28 but for MCD64.

Annex 8 Population estimates of error matrix entries (e_{ij}) and accuracy measures for the sample of Africa 2016, at long and short sampling units

Table 7: Estimated error matrices and reference burned area (m^2) for each product. Standard errors of the estimates are shown in parentheses.

	e_{11}	e_{12}	e_{21}	e_{22}	BA_{ref}
FireCCISFD11 (long unit)	6.58e+12 (2e+12)	1.57e+12 (5e+11)	2.37e+12 (8e+11)	1.14e+14 (6e+13)	8.95e+12 (2e+12)
FireCCIS1A10 (long unit)	1.6e+12 (2e+12)	2.11e+12 (2e+12)	1.83e+12 (1e+12)	1.86e+13 (1e+13)	3.43e+12 (3e+12)
FireCCILT10 (long unit)	2.81e+12 (9e+11)	4e+12 (1e+12)	5.88e+12 (2e+12)	1.26e+14 (6e+13)	8.68e+12 (2e+12)
FireCCI50 (long unit)	2.98e+12 (7e+11)	9.44e+11 (3e+11)	5.71e+12 (2e+12)	1.29e+14 (6e+13)	8.7e+12 (2e+12)
FireCCI51 (long unit)	3.95e+12 (9e+11)	1.37e+12 (4e+11)	4.74e+12 (1e+12)	1.29e+14 (6e+13)	8.68e+12 (2e+12)
MCD64 (long unit)	3.62e+12 (9e+11)	9.29e+11 (2e+11)	5.34e+12 (2e+12)	1.05e+14 (6e+13)	8.96e+12 (2e+12)
FireCCISFD11 (short unit)	2.92e+12 (9e+11)	5.23e+12 (1e+12)	6.03e+12 (1e+12)	1.1e+14 (6e+13)	8.95e+12 (2e+12)
FireCCIS1A10 (short unit)	7.38e+11 (8e+11)	2.98e+12 (3e+12)	2.69e+12 (2e+12)	1.77e+13 (1e+13)	3.43e+12 (3e+12)
FireCCILT10 (short unit)	4.62e+11 (2e+11)	6.34e+12 (2e+12)	8.22e+12 (2e+12)	1.24e+14 (6e+13)	8.68e+12 (2e+12)
FireCCI50 (short unit)	2e+12 (5e+11)	1.92e+12 (5e+11)	6.69e+12 (2e+12)	1.28e+14 (6e+13)	8.7e+12 (2e+12)
FireCCI51 (short unit)	2.58e+12 (6e+11)	2.74e+12 (7e+11)	6.11e+12 (2e+12)	1.27e+14 (6e+13)	8.68e+12 (2e+12)
MCD64 (short unit)	3.02e+12 (7e+11)	1.53e+12 (4e+11)	5.94e+12 (2e+12)	1.04e+14 (6e+13)	8.96e+12 (2e+12)



Table 8: Estimated accuracy of each product. Standard errors of the estimates are shown in parentheses.

	DC	relB	Ce	Oe
FireCCISFD11 (long unit)	0.770 (0.030)	-0.0896 (0.073)	0.193 (0.036)	0.265 (0.047)
FireCCIS1A10 (long unit)	0.448 (0.121)	0.0827 (0.316)	0.569 (0.097)	0.533 (0.176)
FireCCILT10 (long unit)	0.362 (0.043)	-0.216 (0.171)	0.588 (0.058)	0.677 (0.057)
FireCCI50 (long unit)	0.473 (0.056)	-0.548 (0.078)	0.240 (0.029)	0.657 (0.058)
FireCCI51 (long unit)	0.564 (0.041)	-0.388 (0.068)	0.257 (0.028)	0.545 (0.050)
MCD64 (long unit)	0.536 (0.056)	-0.492 (0.070)	0.204 (0.022)	0.596 (0.060)
FireCCISFD11 (short unit)	0.342 (0.038)	-0.0896 (0.073)	0.641 (0.039)	0.674 (0.041)
FireCCIS1A10 (short unit)	0.207 (0.063)	0.0827 (0.316)	0.801 (0.050)	0.785 (0.088)
FireCCILT10 (short unit)	0.0597 (0.012)	-0.216 (0.171)	0.932 (0.014)	0.947 (0.013)
FireCCI50 (short unit)	0.318 (0.043)	-0.548 (0.078)	0.489 (0.044)	0.769 (0.041)
FireCCI51 (short unit)	0.368 (0.039)	-0.388 (0.068)	0.515 (0.034)	0.703 (0.041)
MCD64 (short unit)	0.447 (0.052)	-0.492 (0.070)	0.336 (0.027)	0.663 (0.054)

Annex 9 Accuracy observations at TSAs for the sample of Africa 2016

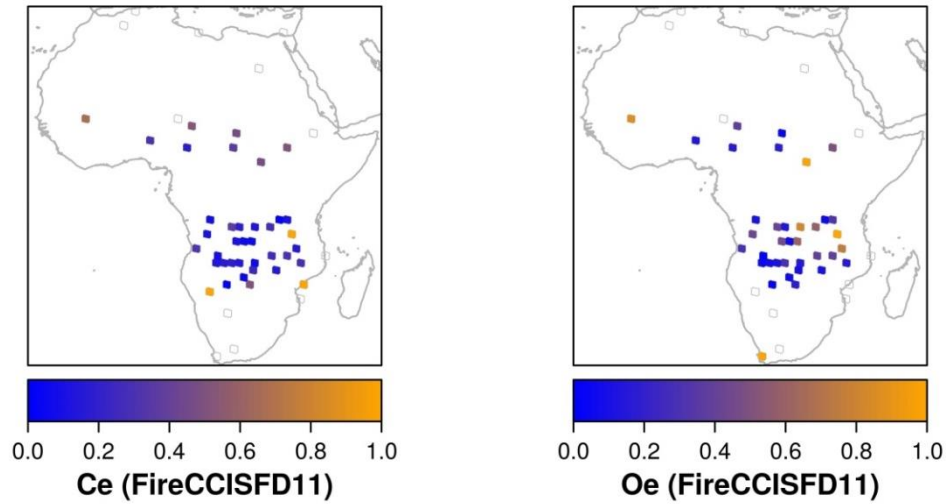


Figure 33: Ce and Oe for FireCCISDF11 at TSAs for long sampling units over the sample of Africa 2016. Units without product data or without accuracy measure available are represented by empty polygons (white polygons with grey borders).

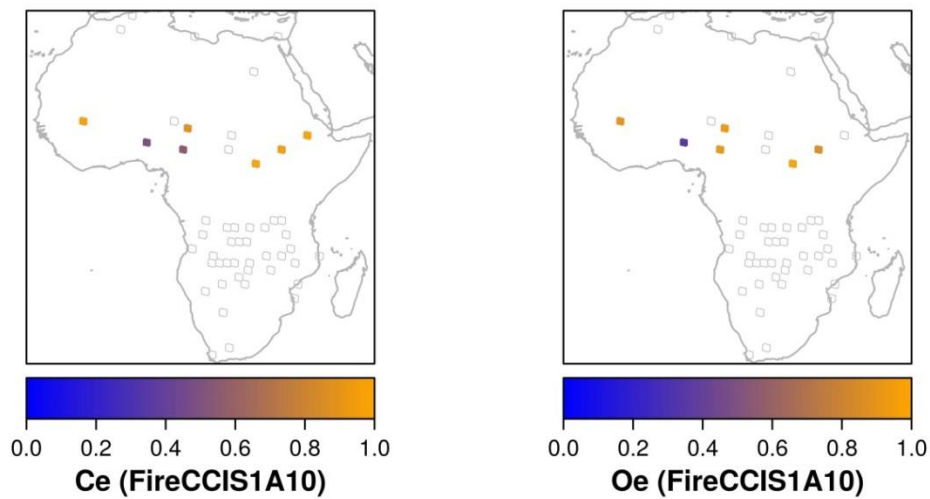


Figure 34: As in Figure 33 but for FireCCIS1A10.

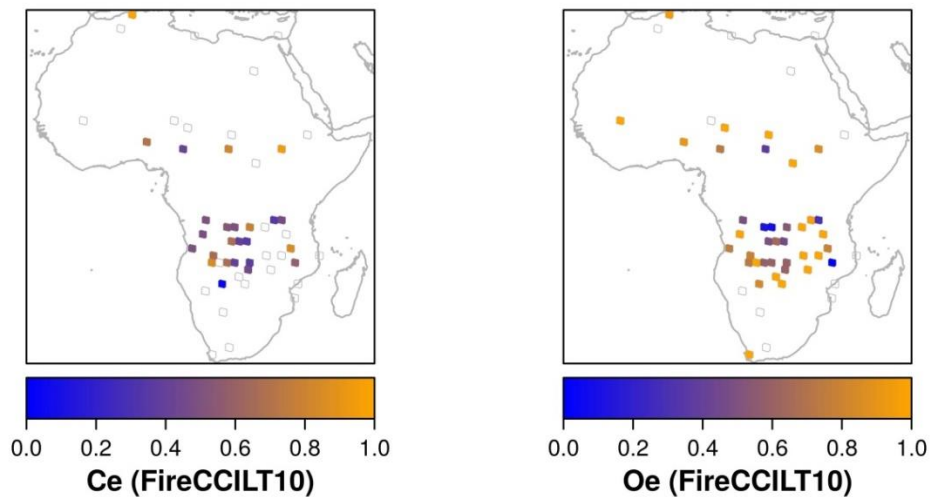


Figure 35: As in Figure 33 but for FireCCILT10.

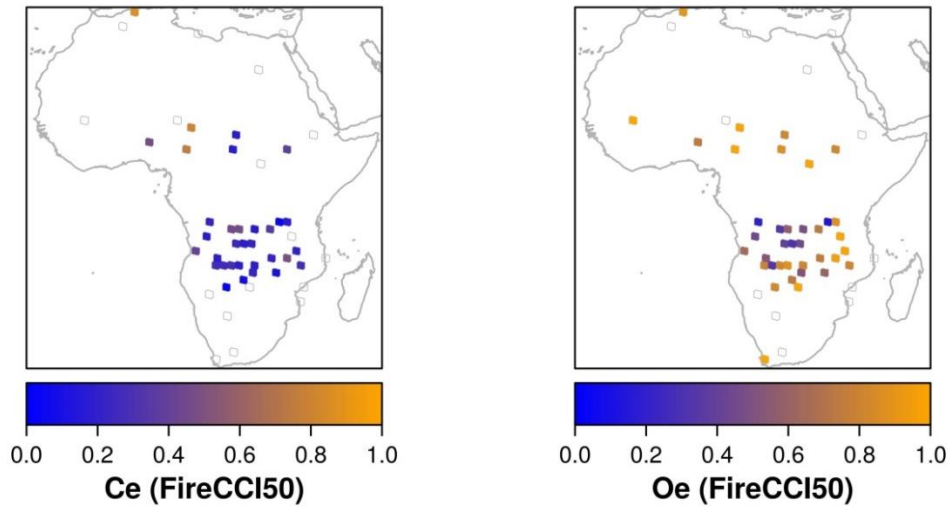


Figure 36: As in Figure 33 but for FireCCI50.

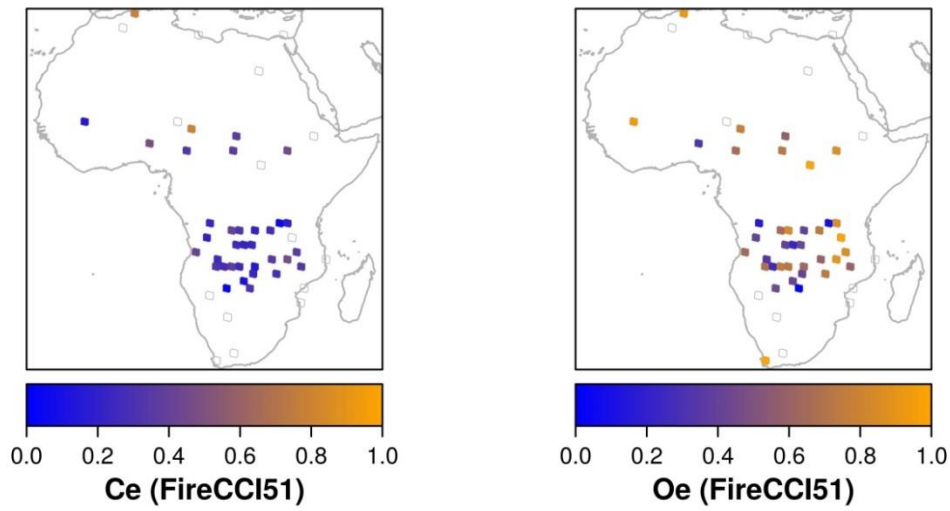


Figure 37: As in Figure 33 but for FireCCI51.

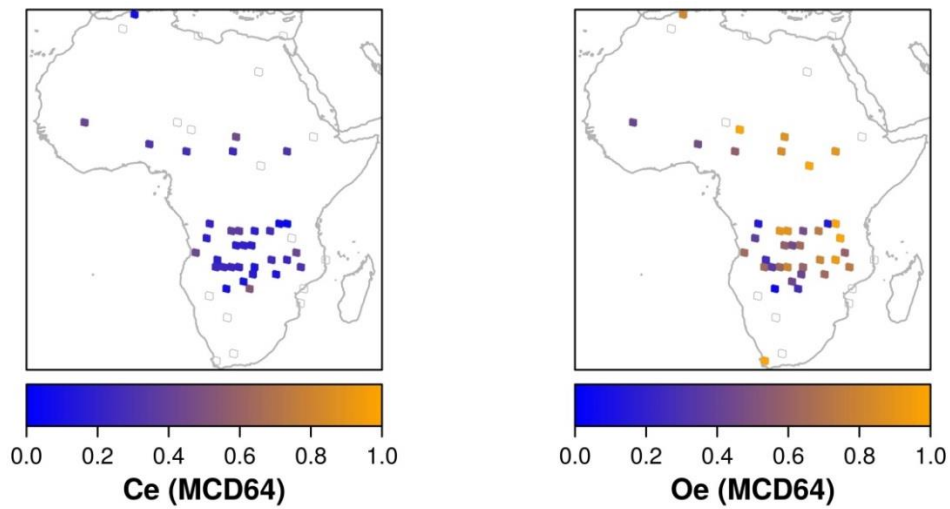


Figure 38: As in Figure 33 but for MCD64.