Elsevier Editorial System(tm) for Remote Sensing of Environment
Manuscript Draft

Manuscript Number: RSE-D-14-00771R2

Title: Comparing the Accuracies of Remote Sensing Global Burned Area Products using Stratified Random Sampling and Estimation

Article Type: Original Research Paper

Keywords: Validation, Error matrix, Probability sampling, Fire Disturbance

Corresponding Author: Dr. Marc Padilla,

Corresponding Author's Institution: University of Alcalá

First Author: Marc Padilla

Order of Authors: Marc Padilla; Stephen V. Stehman; Ruben Ramo; Dante Corti; Stijn Hantson; Patricia Oliva; Alonso-Canas Itziar; Andrew V. Bradley; Kevin Tansey; Bernardo Mota; Jose Miguel Pereira; Emilio Chuvieco

Highlights (for review)

- Statistical methods were applied to compare the accuracy of global burned area products.
- Probability sampling was used to infer product accuracy for the globe.
- MODIS MCD64 was the most accurate burned area product evaluated.

Comparing the Accuracies of Remote Sensing Global Burned Area Products using Stratified Random Sampling and Estimation

Marc Padilla[ac]*, Stephen V. Stehman[b], Ruben Ramo[a], Dante Corti[a], Stijn Hantson[a], Patricia Oliva[a], Itziar Alonso-Canas[a], Andrew V. Bradley[cd], Kevin Tansey[c], Bernardo Mota[e,f], Jose Miguel Pereira[e] and Emilio Chuvieco[a]

a Environmental Remote Sensing Research Group, Universidad de Alcalá, Colegios 2, Alcalá de Henares, Spain .

b Department of Forest and Natural Resources Management, College of Environmental Science and Forestry, State University of New York, Syracuse, NY, 13210, USA.

c Department of Geography, University of Leicester, Leicester, LE1 7RH. United Kingdom.

d Imperial College of London, Silwood Park Campus, SL5 7YP. United Kingdom.

e Forest Research Centre, School of Agriculture, Technical University of Lisbon, 1349-017 Tapada da Ajuda St., Lisbon, Portugal.

f Department of Geography, Kings College London, Strand, London WC2R 2LS. United Kingdom.

E-mail addresses: mp489@le.ac.uk (M. Padilla), svstehma@syr.edu (S.V. Stehman), dcorti@infor.cl (Dante Corti), hantson.stijn@gmail.com (S. Hantson), patricia.oliva@uah.es (P. Oliva), itziar.alonsoc@uah.es (I. Alonso), a.bradley@imperial.ac.uk (A. Bradley), kjt7@leicester.ac.uk (K. Tansey), bernardo.mota@kcl.ac.uk (B. Mota), jmcpereira@isa.utl.pt (J.M. Pereira), emilio.chuvieco@uah.es (E. Chuvieco).

*Corresponding author. Tel.: +34 918854482; fax: +34 918854439

Abstract

The accuracies of six global burned area (BA) products for year 2008 were compared using the same validation methods and reference data to quantify accuracy of each product. The selected products include MCD64, MCD45 and Geoland2, and three products developed within

the Fire Disturbance project (fire_cci), which is part of the European Space Agency's (ESA) Climate Change Initiative (CCI) program. The latter three products were derived from MERIS and VEGETATION sensors (one product from each sensor separately, and a third one from the merging of MERIS and VGT products). The reference fire perimeters were mapped from two multi-temporal Landsat TM/ETM+ images at 103 non-overlapping Thiessen scene areas (TSA) selected with a stratified random sampling design. The validation results were based on cross tabulated error matrices from which six accuracy measures were computed following the requirements of end-users of burned area products. While overall accuracy (OA) exceeded 99% for all products, overall accuracy was lower for the burned class. Burned area commission error ratio was above 40% for all products and omission error ratio was above 65% for all products. The statistical significance of differences in accuracy between pairs of products was evaluated based on theory of the stratified combined ratio estimator. Statistical tests identified the MCD64 as the most accurate product, followed by MCD45 and the MERIS product.

Keywords: Validation, Error matrix, Probability sampling, Fire Disturbance.

1. Introduction

Fire affects atmospheric emissions of gases and aerosols (van der Werf et al. 2004) and influences carbon budgets, as it impacts carbon stocks and vegetation succession patterns. Therefore, accurate information on fire occurrence is critical to better understand the role of vegetation dynamics in earth system models (Bowman et al. 2009). For this reason, the Global Climate Observing System (GCOS) program (GCOS 2004) identified Fire disturbance as one of the Essential Climate Variables (ECV). This variable has been selected by the European Space Agency (ESA) as one of the target variables for the Climate Change Initiative (CCI) program (http://ionia1.esrin.esa.int/, last accessed December, 7th 2014). The ESA CCI Fire Disturbance project (fire_cci) aimed to develop global burned area products from European sensors for the

50 climate modeling community, with proper validation and uncertainty characterization
51 (http://www.esa-fire-cci.org/, last accessed December, 7th 2014).

52 In the last few years, several global burned area (BA) products have been made available to
53 the international community, and are being used as input to climate models (Mouillot et al.
54 2014). Independent validation assessments are necessary to compare the performance of
55 these products and guide their use when incorporated into global atmospheric and carbon
56 models. Knowing the uncertainty of each input product is critical to decouple model and input
57 data limitations.

58 Validation is defined by The Committee on Earth Observing Satellites Working Group on
59 Calibration and Validation (CEOS-WGCV) as "the process of assessing, by independent means,
60 the quality of the data products derived from the system outputs" (CEOS-WGCV 2012).
61 Validation quantitatively assesses the performance of a dataset providing essential
62 information to the user community. Existing BA products have typically been subject to a first
63 stage validation. Globcarbon (Plummer et al. 2007) and L3JRC (Tansey et al. 2008) were
64 validated with independent data derived from 72 globally distributed Landsat scenes mostly
65 acquired from the year 2000. Chuvieco et al. (2008) validated a regional product for Latin
66 America using 19 Landsat scenes and 9 China–Brazil Earth Resources Satellite (CBERS) scenes.
67 Roy and Boschetti (2009) reported validation results for the MCD45 (Roy et al. 2008) product
68 and Giglio et al. (2009) for MCD64, the former in southern Africa using 11 Landsat scenes and
69 the latter using 41 Landsat scenes in the western United States, southern Africa and central
70 Siberia.

71 Extending the objective to comparing the accuracy of different global products is still a
72 challenge as validation methods and datasets are not fully compatible. Roy and Boschetti
73 (2009) presented a first attempt at the validation and comparison of several global products

using a common independent reference data set. They compared Globcarbon, MCD45 and L3JRC BA products with fire perimeters derived from 11 Landsat scenes distributed across southern Africa. Results were reported for each Landsat scene, but global accuracy for the whole study area was not reported.

In this paper we compare the accuracy of six burned area products at a global scale for year 2008 using a stratified random sample developed for the fire_cci project, which was the first attempt to implement a statistically designed sample for global validation of burned area products (Padilla et al. 2014a). The accuracy measures used to compare the burned area products were selected to address the requirements defined by the end-users of the fire_cci products (Mouillot et al. 2014). Specifically, users expressed interest in metrics providing estimates of accuracy, commission and omission errors, error bias (whether the product under or overestimates true BA) and temporal stability (covered in Padilla et al. 2014b). We also employed a statistical test to evaluate the differences in accuracy of product pairs.

2. Methods

2.1. BA products

The products evaluated in this study (Table 1) include two products derived from MODIS (Moderate Resolution Imaging Spectroradiometer), MCD64 and MCD45, one developed in the Geoland2 project from SPOT VEGETATION (VGT) data, and three products developed in the fire_cci project.

The fire_cci project has generated three BA products: the first one derived from SPOT VGT (VGT_cci) and based on a time series change detection algorithm to detect significant decreases in the near-infrared reflectance (Pereira et al. 2013), another one computed from the MEdium Resolution Imaging Spectrometer (MERIS), (MERIS_cci), which used an hybrid algorithm that takes into account both reflectance time series and MODIS active fire

observations (Alonso-Canas and Chuvieco, submitted), and a third product (MERGED_cci) based on the merging of VGT and MERIS data (Tansey et al. 2014).

MCD45 is currently the standard MODIS BA product. It is based on a prognostic model that compares estimated versus actual reflectance for different MODIS spectral bands (Roy et al. 2005). MCD64 is the primary data source of the Global Fire Emissions Database (GFED) Versions 3 and 4 (Giglio et al. 2010; 2013). This product is based on MODIS spectral indices and active fire observations (Giglio et al. 2009). Geoland2, based on a temporal index of near infrared reflectances of the SPOT VGT sensor, is built on the experiences of the Global Burned Area (GBA2000), Globcarbon and L3JRC projects (Tansey et al. 2012; 2008). The latter two could not be assessed in this paper, as they do not include 2008 data, which was selected as the golden validation year for the ESA CCI program.

All six BA products compared in this study include monthly files with pixel values referring to the day of the year (DoY) when a burned area was detected (1-365, 0 meaning unburned).

Table 1: List of products included in the analysis

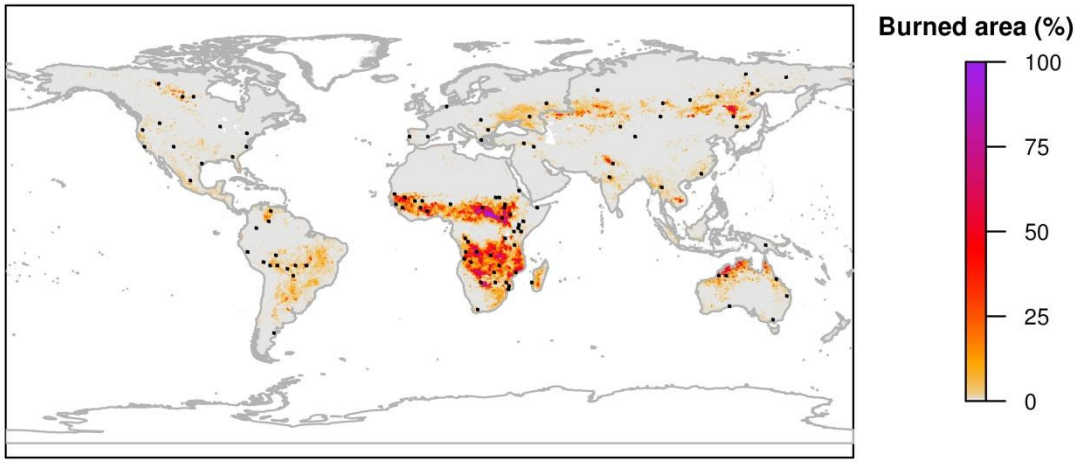
Acronym	Sensor characteristics	Project	Institution
MCD45	MODIS images (500m)	MCD45 (Roy et al. 2005)	University of Maryland
MCD64	MODIS images (500m), MODIS thermal anomalies (1km)	MCD64 (Giglio et al. 2009)	University of Maryland
Geoland2	SPOT VGT (1km)	Geoland2 (Tansey et al. 2012; 2008)	University of Leicester and Flemish Institute for Technological Research (VITO)

MERGED_cci	SPOT VGT and MERIS (300m)	Fire Disturbance CCI project (Chuvieco 2013)	University of Leicester
MERIS_cci	MERIS (300m) MODIS thermal anomalies (1km)		University of Alcalá
VGT_cci	SPOT VGT (1km)		Instituto Superior de Agronomia

112

113 **2.2. Sampling Design**

114 The sampling design, reference data generation and methodology for estimating accuracy had
115 previously been documented in Padilla et al. (2014a), where further details are included. The
116 probability sampling design employed a spatial stratification to distribute the sample among
117 the major Olson biomes (Olson et al. 2001), with proportionally larger sample sizes allocated to
118 regions with high BA. Two levels of stratification were implemented using, as sampling units,
119 the Thiessen scene areas (TSAs) constructed by Cohen et al. (2010) and Kennedy et al. (2010)
120 specifically for use with Landsat WRS-II frames. The first stratification level was based on the
121 Olson biomes and the second one on the BA extent in 2008 provided by the Global Fire
122 Emissions Database (GFED) version 3 (Giglio et al. 2009; Giglio et al. 2010). Fourteen strata
123 were defined; each one of the seven biome-based (geographic) strata was split into two
124 regions of high and low BA. The global distribution of the sample is illustrated in Figure 1.
125 Globally, 103 TSAs were analyzed out of the 105 selected for the sample. Two TSAs were
126 excluded because at least one of the BA products did not report results for the region within
127 which that TSA was located. Specifically, the MCD64 had all pixels with no-data available in one
128 TSA, and MCD45 in a second TSA.



129

130 **Figure 1: Spatial distribution of the 103 sample TSAs (black polygons) selected by stratified random**
131 **sampling (for context, % burned area in 2008 at 0.5° spatial resolution is shown based on GFED**
132 **version 3 (Giglio et al. 2010)).**

133

134 ### 2.3. Reference data

135 The standard protocol defined by the CEOS Cal-Val (Boschetti et al. 2009) was followed to
136 generate and document the fire reference perimeters for 2008, the year selected for validation
137 of all ESA CCI products. For each TSA sampled, fire perimeters were extracted from a pair of
138 Landsat TM/ETM+ image acquisitions at the same location (acquired in two different revisit
139 times at the same path and row), using a semi-automatic algorithm developed by Bastarrika et
140 al. (2011). All scenes were afterwards visually checked and some were repeated by another
141 interpreter to ensure consistency of the results (see Padilla et al., 2014a).

142 ### 2.4. Accuracy measures

143 The validation results are based on the cross tabulation or error matrix approach (Table 2) to
144 summarizing accuracy (Congalton and Green 1999; Latifovic and Olthof 2004). Parameter
145 estimates of the error matrix ($\hat{P}_{ij}$) are obtained from the sampled pixels with available product

and reference data. To build the error matrices, product pixels were coded as "burned" if fire was detected between the reference image acquisition dates. The rest of the product pixels were coded as "unburned" or "no-data", the latter for unobserved pixels.

The error matrix cell entries for pixel *u* ($e_{ij,u}$) are based on the proportion of area of agreement or disagreement in that pixel (Padilla et al. 2014a). The error matrix of a TSA *t* is based on the sum of its single pixel error matrices ($E_{ij,t} = \sum_{u \in t} e_{ij,u}$; the summation is over all interpretable pixels within TSA *t,* N_t). Error matrices can be expressed in terms of proportion of area if divided by the number of interpretable pixels. The population error matrix expressed as proportion of area is based on the sum of e_{ij} values, $E_{ij} = \sum_t E_{ij,t}$, divided by the number of interpretable pixels (*N*) for the region of interest, $P_{ij} = E_{ij} / N$.

Table 2: Population error matrix where P_{ij} is expressed as proportion of area of agreement (diagonal cells) or disagreement (off diagonal cells) between the BA product (map) class and the reference class ($\sum P_{ij}$=1, and the row and column margins are the sum of the cell entries in that row or column).

	Reference data		
Product prediction	Burned	Unburned	Row total
Burned	P_{11}	P_{12}	P_{1+}
Unburned	P_{21}	P_{22}	P_{2+}
Col. Total	P_{+1}	P_{+2}	*1*

The computed accuracy measures are overall accuracy,

161 $$OA = \frac{E_{11} + E_{22}}{N} \quad (1)$$

162 commission error ratio,

163 $$Ce = E_{12} / E_{1+} \quad (2)$$

164 and omission error ratio,

165 $$Oe = E_{21} / E_{+1} \quad (3)$$

166 the two latter referring to the category “burned” (Boschetti et al. 2004; Chuvieco et al. 2008;
167 Roy and Boschetti 2009). A comparison of products based on those two burned accuracy
168 measures may not yield an unambiguously preferred product. Particularly, when the user
169 regards BA omission and commission errors as equally important and when they vary in
170 different magnitudes and directions (e.g. Ce increases and Oe decreases) it is difficult to decide
171 which product is preferred in terms of the category “burned” accuracy. Therefore we
172 additionally report the Dice coefficient (DC), which summarizes both errors in a single metric:

173 $$DC = \frac{2E_{11}}{2E_{11} + E_{12} + E_{21}} = \frac{2E_{11}}{E_{1+} + E_{+1}} \quad (4)$$

174 DC has a sensible probabilistic interpretation (Dice 1945; Fleiss 1981; Forbes 1995; Hand 1981;
175 Hellden 1980; Liu et al. 2007) which is the following: Given that one classifier (product or
176 reference data in our case) identifies a burned pixel, DC is equal to the conditional probability
177 that the other classifier will also identify it as burned (Fleiss 1981).

178 Two measures of bias were computed, as the users of the fire_cci project reported their
179 interest in having a BA product with error balance (Mouillot et al. 2014). Bias is defined in
180 terms of proportion of BA:

181 $$B = \frac{E_{1+} - E_{+1}}{N} = \frac{E_{12} - E_{21}}{N} \quad (5)$$

182 Relative bias (relB), which is scaled to the reference BA,

183 $$relB = \frac{E_{1+} - E_{+1}}{E_{+1}} = \frac{E_{12} - E_{21}}{E_{+1}} \quad (6)$$

184 The sign of B and relB values indicates whether a product overestimates (positive sign) or
185 underestimates the extent of BA (negative sign).

186

187 ### 2.5. Global description of accuracy

188 Global error matrices and the derived accuracy measures were estimated taking into account
189 the stratified sampling design (Padilla et al. 2014a; Stehman et al. 2007). The general estimator
190 for an accuracy measure is a stratified combined ratio estimator (Cochran 1977) of the form

191 $$\hat{R} = \frac{\sum_{h=1}^{H} K_h \bar{y}_h}{\sum_{h=1}^{H} K_h \bar{x}_h} \quad (7)$$

192 where H is the number of strata, K_h is the size of stratum h, $\bar{y}_h$ and $\bar{x}_h$ are the sample means
193 of y_t and x_t at stratum h, and x_t and y_t are values defined by the denominator and numerator of
194 the different accuracy measures at TSA t (Padilla et al. 2014a; Appendix A). The estimated
195 variance of $\hat{R}$ is

196 $$\hat{V}(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h=1}^{H} \frac{K_h^2}{k_h(k_h - 1)} \sum_{t \in h} d_t^2 \quad (8)$$

where $\hat{X} = \sum_{h=1}^{H} K_h x_h$, $d_t = (y_t - \bar{y}_h) - \hat{R}(x_t - \bar{x}_h)$ and k_h is the number of TSAs sampled in stratum *h*. The standard error of $\hat{R}$ is the square root of $\hat{V}(\hat{R})$.

To assess the utility of the stratification, we evaluated the improvement in precision achieved by the stratified design relative to a proportionally allocated stratified sample, following the approach for ratio estimates of Cochran (1977; Section 6.14). It is assumed that an improvement in the precision relative to a proportional sample indicates a similar or larger improvement relative to a simple random sample (Cochran 1977; Section 5.7), given that the variances within strata are not expected to be higher than the variances among strata. The ratio of the variance for the stratified design implemented relative to the variance for a proportionally allocated stratified design provided a quantitative assessment of the utility of the stratified design used in this study.

2.6. Comparison between product accuracies

The difference between the accuracy measures of two products was assessed with the difference between the combined ratio estimators, $\hat{R} - \hat{R}'$ (for example, if the comparison is based on omission error, $\hat{R}$ would be Oe for one BA product and $\hat{R}'$ would be Oe for the other BA product). The estimated variance of $\hat{R} - \hat{R}'$ was computed taking into account the stratified sample design and the possible correlation between the two estimated ratios (Cochran 1977).

$$\hat{V}(\hat{R} - \hat{R}') = \hat{V}(\hat{R}) + \hat{V}(\hat{R}') - 2\,\hat{\mathrm{cov}}(\hat{R}, \hat{R}') \quad (9)$$

where

$$\hat{\mathrm{cov}}(\hat{R}, \hat{R}') = \frac{1}{\hat{X}\hat{X}'} \sum_{h=1}^{H} \frac{K_h^2}{k_h(k_h - 1)} \sum_{t \in h} d_t d_t' \quad (10)$$

The denominators of equations (8) and (10) are based on total estimates ($\hat{X}$ and $\hat{X}'$) instead of mean estimates as specified in Cochran (1977; equations 9.92 and 9.93). This modification allows for exact equivalent results with the formulas of Cochran (1977; Section 6.11).

Typically the estimated covariance (equation 15) will be positive and may lead to a substantial reduction in the estimated variance (equation 13) relative to the case of two independent samples. The statistical significance of the difference in accuracy between two products (i.e., the difference between the parameters *R* and *R*') was evaluated with the z statistic

$$z = \frac{\hat{R} - \hat{R}'}{\sqrt{\hat{V}(\hat{R} - \hat{R}')}} \qquad (11)$$

which follows a standard normal distribution (assuming a large sample size). To determine whether one product has different accuracy from another product the null hypothesis of no difference would be rejected at the 0.003 (0.05/15) level of significance, which is equivalent to the widely used 5 percent level once the multiple testing feature is taken into account using the Bonferroni method (Miller 1966). That is, the Bonferroni method ensures that for the set of 15 pairwise comparisons among the six BA products for a given accuracy metric, the probability of at least one Type I error is 0.05 or smaller.

Because low bias (regardless of the sign) is a desirable feature of a BA produce (i.e., minimal overestimation or underestimation of BA), the comparison of bias required an additional step not present in other comparisons. Specifically, for a pair of products with different averaged bias signs, the negative bias was re-defined as P_{21}-P_{12}. For example, if B=0.03 for one BA product and B=-0.01 for another, instead of using the simple difference (0.04), we used the actual deviation from 0 (0.03-0.01=0.02), and whether this difference would be statistically significant. This modification made it possible to know which product was closest to the lack of bias (B=0 and relB=0).

3. Results

The estimated error matrices ($\hat{P}_{ij}$) and accuracy measures describing accuracy for the globe are presented in Tables 3-4. Standard errors of accuracy estimates were relatively small and similar for all products (e.g. standard error of DC smaller than 5% in all products). Further confirmation of the utility of the stratified design was evidenced by the assessments of the gains in precision. Had simple random sampling been implemented with the same sample size, the variance for estimating the accuracy measures would have been at least 1.75 times the variance obtained from the stratified design in all cases with the exception of OA, B and relB in VGT_cci and MERGED_cci.

Table 3: Estimated error matrices and reference burned area percent (BA%) for each product at the global level. Cells of the error matrices are represented by estimated percent area, 100* $\hat{P}_{ij}$. $\hat{P}_{11}$ represents the estimated proportion of area that is burned according to both the BA product and the reference classification, $\hat{P}_{12}$ is the estimated proportion of BA commission error, $\hat{P}_{21}$ is the estimated proportion of BA omission error, and $\hat{P}_{22}$ is the estimated proportion of area that is unburned according to both the BA product and the reference classification.

Product	$\hat{P}_{11}$	$\hat{P}_{12}$	$\hat{P}_{21}$	$\hat{P}_{22}$	BA%
MCD64	0.13	0.10	0.27	99.5	0.40
MCD45	0.10	0.08	0.25	99.6	0.34
Geoland2	0.03	0.08	0.32	99.6	0.35
MERGED_cci	0.07	0.49	0.27	99.2	0.35
MERIS_cci	0.08	0.15	0.27	99.5	0.35
VGT_cci	0.03	0.40	0.38	99.2	0.41

Table 4: Estimated accuracy (expressed as %) of each product at the global level. Standard errors are shown in parentheses and superscript letters refer to products that have statistically significantly (α=0.003 per comparison) higher accuracy or lower bias than the designated product according to z tests. For example, the f superscript for MCD64 on the DC measure indicates that MCD64 has significantly higher DC than product f=VGT_cci. The test results for OA are not shown because OA differs very little among all products.

Product	OA(%)	Ce(%)	Oe(%)	DC(%)	B(%)	relB(%)
a MCD64	99.6 (0.1)	42 (4)^{cdef}	68 (4)^{cdef}	42 (4)^{cdef}	-0.17 (0.05)	-44 (7)
b MCD45	99.7 (0.1)	46 (4)^{cdef}	72 (5)^{cf}	37 (4)^{cdf}	-0.17 (0.05)	-48 (8)
c Geoland2	99.6 (0.1)	74 (9)^{f}	91 (4)	13 (5)	-0.23 (0.07)	-68 (9)
d MERGED_cci	99.2 (0.2)	87 (5)^{f}	79 (5)^{cf}	16 (5)^{f}	0.21 (0.18)	62 (55)
e MERIS_cci	99.6 (0.1)	64 (6)^{df}	76 (5)^{cf}	29 (5)^{cdf}	-0.12 (0.04)	-34 (12)
f VGT_cci	99.2 (0.2)	94 (4)	93 (2)	7 (3)	0.02 (0.22)	4 (54)^{d}

All products were validated with a common set of reference sites; however, VGT_cci and MCD64 had slightly higher BA% (>0.40%) relative to the other four products (see right column of Table 3). This may be attributable to VGT_cci and MCD64 having more no-data pixels within the reference unburned area than in the reference burned area. The estimated error matrices and accuracy measures considering no-data of products as unburned are presented in the Appendix, for all 105 TSAs originally selected. Minor differences were observed in the results with respect to using 103 sites where all products have valid data. Greater differences among BA products were observed in the quantity of available data. The proportion of available pixels

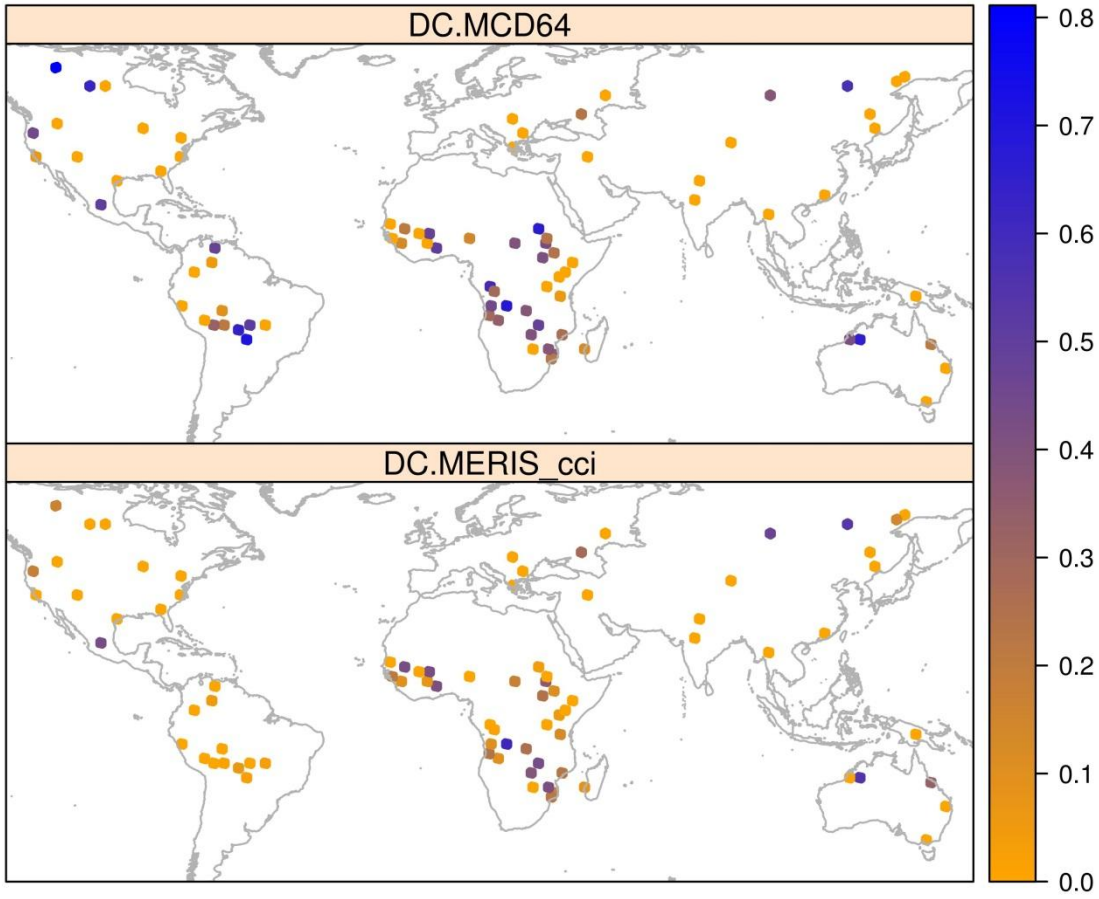
272 in the 105 TSAs was near 82% for VGT_cci and MCD64, 97% for MCD45, and 100% for the
273 other products.

274 While overall accuracy (OA) exceeds 99% for all products, the accuracy results are less
275 favorable for the burned class. Burned area commission error ratio (Ce) is above 40% for all
276 products and above 90% for VGT_cci . Burned area omission error ratio (Oe) is above 65% for
277 all products and above 90% for VGT_cci and Geoland2.. The Oe translates to underestimation
278 of BA, except for VGT_cci and MERGED_cci. The latter product is the only one that
279 overestimates BA (estimated bias of 0.21% in terms of percent of area, and 62% in terms of
280 relative bias). For MCD64, MCD45, Geoland2 and MERIS_cci, the bias (B) ranged from -0.23%
281 to -0.12% in terms of percent of area, and relative bias (relB) ranged from -68% to -34%.

282 The pairwise comparison of product accuracies using the accuracy measures and the z test
283 identified MCD64 as the best performing product. It had statistically significantly (α=0.003)
284 higher accuracy than all other products except for MCD45 on DC, Ce and Oe. MCD45 was the
285 second best based on the accuracy scores. It had significantly higher DC than all other products
286 except MCD64 and MERIS_cci. MERIS_cci was the third best performing product as it had
287 higher accuracy than the other two fire_cci products and Geoland2, on DC and in Ce and Oe.
288 Figure 2 shows the spatial distributions of accuracy for the best performing product (MCD64)
289 and MERIS_cci. Some TSAs mainly located in Africa, tended to have the lowest values of Ce and
290 Oe and the highest values of DC, for MCD64 and MERIS_cci. DC values were not available for
291 some TSAs (TSAs missing in the DC panels of figures below), as they cannot be computed when
292 BA data in either the reference or the global product are absent. Nevertheless, as it is
293 described in Section 2.5 those matrices contributed as well to the global error matrices.

294 Four products had large underestimation of BA while MERGED_cci and VGT_cci had large and
295 moderate overestimation respectively (Table 4). The low estimated bias with a large variance

296 of VGT_cci (relB was 4% and its standard error 54%) reflects that the per-TSA overestimates
297 and underestimates offset yielding a small final overall bias. Figure 3 shows that for the
298 MERGED_cci most TSAs have little bias (grey), some TSAs underestimate BA (red) and others
299 overestimate BA (green). In contrast, for MCD64 and MCD45 almost all TSAs have small bias
300 (grey) or underestimation (red), as can be seen in Figure 3 for MCD64.



301

302 **Figure 2: Accuracy measures of DC on TSAs for MCD64 and MERIS_cci (TSA spatial extents are**
303 **exaggerated to enhance visualization).**

304

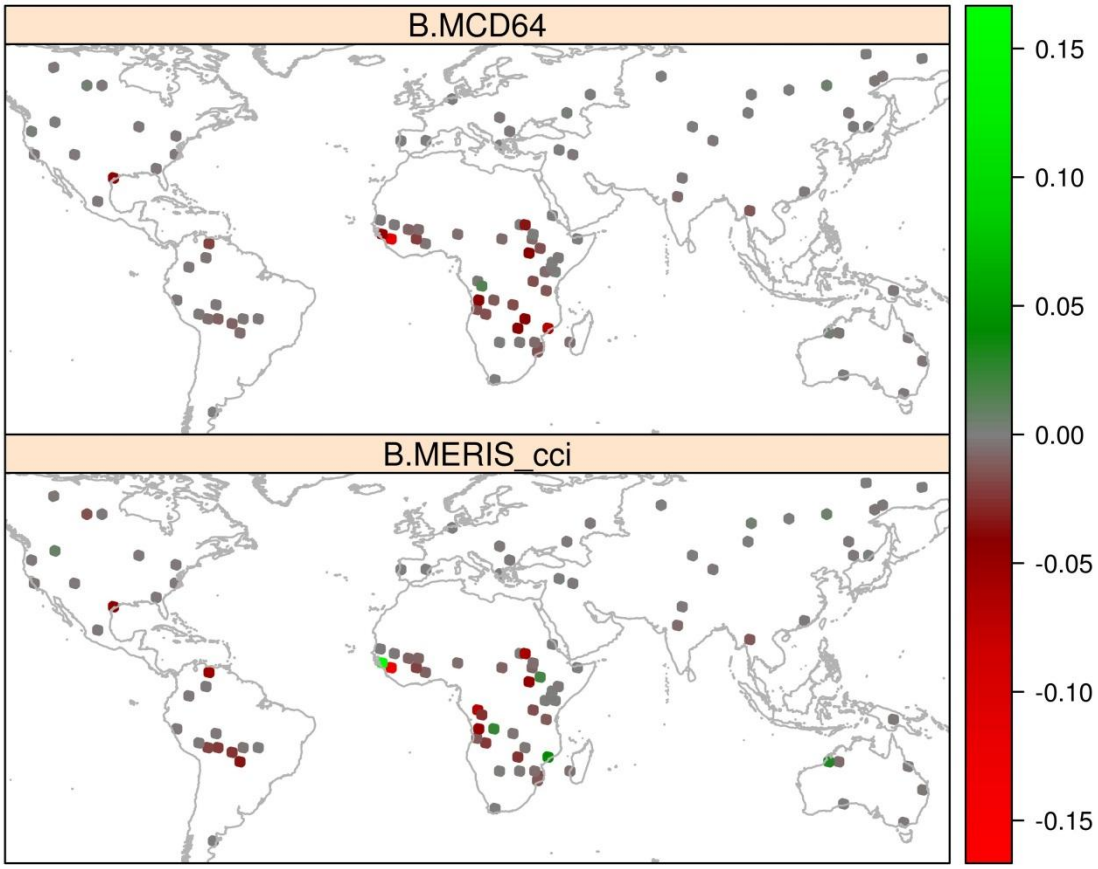


Figure 3: Accuracy measures of B on TSAs for MCD64 and MERIS_cci (TSA spatial extents are exaggerated to enhance visualization).

4. Discussion

In this paper, six BA products were compared using error matrices and accuracy measures derived from a common reference sample, selected following a probability (stratified random) sampling design. Statistical significance of differences in accuracy measures among BA products was evaluated taking into account the probability sampling protocol. Although the methodology described provides techniques to compare BA products, the comparison does not establish whether the differences between product accuracies have practical importance, nor does the comparison establish whether the product is sufficiently accurate for a user's

317 intended application. Even the most accurate product identified by the comparison is not
318 guaranteed to meet a user's needs, and conversely, the least accurate product may still be
319 adequate for certain user's objectives.

320 The statistical tests identified MCD64 as the most accurate product followed by MCD45 and
321 MERIS_cci. The very high values of OA (larger than 99% for all products) and its lack of
322 variability among products reflect the problems of the OA metric when one of the categories
323 has a strong prevalence over the other (Fielding and Bell 1997). Within a region where very
324 little area is burned (which is the common case) any classification algorithm even with low
325 accuracy in the burned category but with a modest level of accuracy in the unburned category
326 would still provide high values of OA. Medium to high underestimation of BA extent was
327 observed for four products MCD64, MCD45, Geoland2 and MERIS_cci. Roy and Boschetti
328 (2009) reported a similar underestimation for three BA products, one of which was MCD45, in
329 their analysis on southern Africa.

330 The common large underestimation of BA extent may be caused by the coarse spatial
331 resolution of global products and the large amount of area burned in small fire patch sizes. It is
332 very unlikely to detect any burned patch smaller than 4-10 product pixels. For instance Giglio
333 et al. (2009) consider than their MCD64 product would unlikely detect any fires smaller than
334 120 ha. The reference files used for this validation exercise were produced from 0.09 ha pixels
335 (from Landsat-TM or ETM+), and therefore included a large proportion of BA in much smaller
336 patch sizes.

337 Another relevant source of potential errors is the temporal reporting accuracy (Boschetti et al.
338 2010), which is closely linked to both the temporal resolution of each sensor and the
339 cloudiness of each region. This implies that pixels correctly detected as burned but dated after
340 the acquisition date of the second multitemporal Landsat image will be considered as omission

errors. Conversely, fires occurring before the first Landsat image but dated afterwards may be considered as commission errors. We have not checked the relevance of this source of error for all sensors, but it may be particularly important for the MERIS_cci product, as MERIS is the only sensor from those compared that does not have daily acquisition frequency (3 days at the Equator with the Full resolution Mode). Estimated temporal reporting accuracy of this sensor is between 5 and 20 days (Alonso-Canas and Chuvieco, submitted) and therefore it will mainly affect estimated global accuracy in those regions where the acquisition dates of the Landsat images were closer in time (<32 days), mostly in Tropical regions.

An important feature of the validation and comparison of BA products based on cross-tabulation analysis is that the comparison of BA products can be based on accuracy metrics selected to address specific end-user requirements. For example, users primarily interested in the estimate of area burned for a specific region would focus on the criterion of bias (i.e., are there systematic over- or under-estimates of BA?), whereas users requiring minimal BA omission error would focus on that criterion.

The concurrence of high accuracy values at some TSAs, mainly for MCD64, MCD45 and MERIS_cci, suggests that there is a site effect at TSA level, i.e. the accuracy not only depends on the product but also on the site. Multiple causes may be responsible for the observed site effect. For example, specific land cover types, phenology states or landscape characteristics may affect the prediction capacity of BA classification algorithms. Fire patch characteristics can also influence algorithm accuracies. For example, small and irregular BA patches may lead to high omission errors because the spatial resolution of the pixels is too coarse to detect the smaller fires (Boschetti et al. 2004). Conversely, small and irregular unburned areas within large burned patches may lead to an increase in commission errors, particularly for high sensitivity algorithms.

5. Conclusions

A statistically rigorous validation protocol based on a probability sampling design was used to objectively compare the accuracy of six BA products MCD64, MCD45, Geoland2, and three products created as part of the fire_cci project. Product comparisons were based on six measures (OA, Ce, Oe, DC, B and relB) selected to address end-user specified desirable accuracy reporting requirements. For the burned category, MCD64 was the most accurate product followed by MCD45 and MERIS_cci. In concordance with previous studies (Roy and Boschetti 2009), BA extent tended to be underestimated by four products. MERGED_cci and VGT_cci had high overestimation and low bias respectively, although the latter with high variability at the TSA level. Landscape and fire patch characteristics are likely to affect similarly the classification algorithm performance given the observed site effects.

Acknowledgments

The authors would like to thank the European Space Agency for funding the Fire Disturbance project through the Climate Change initiative. A.V. Bradley was supported in part by FP7-ERC grant number 281986.

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., Flasse, S.P., & Brivio, P.A. (2004). Analysis of the conflict between omission and commission in low spatial resolution dichotomic thematic products: The Pareto Boundary. *Remote Sensing of Environment, 91*, 280-292

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 CEOS-CalVal (Ed.) (pp. 1-11). USA: Committee on Earth Observation Satellites

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

Bowman, D.M.J.S., Balch, J.K., Artaxo, P., Bond, W.J., Carlson, J.M., Cochrane, M.A., D'Antonio, C.M., DeFries, R.S., Doyle, J.C., Harrison, S.P., Johnston, F.H., Keeley, J.E., Krawchuk, M.A., Kull, C.A., Marston, J.B., Moritz, M.A., Prentice, I.C., Roos, C.I., Scott, A.C., Swetnam, T.W., Van der Werf, G.R., & Pyne, S.J. (2009). Fire in the Earth System. *Science, 324*, 481-484

CEOS-WGCV (2012). Working Group on Calibration and Validation - Land Product Validation Subgroup. In. http://lpvs.gsfc.nasa.gov/
Chuvieco, E. (2013). ESA CCI ECV Fire Disturbance - Product Specification Document. In: ESA Fire-CCI project
Chuvieco, E., Opazo, S., Sione, W., Del Valle, H., Anaya, J., Di Bella, C., Cruz, I., Manzo, L., López, G., Mari, N., González-Alonso, F., Morelli, F., Setzer, A., Csiszar, I., Kanpandegi, J.A., Bastarrika, A., & Libonati, R. (2008). Global Burned Land Estimation in Latin America using MODIS Composite Data. *Ecological Applications, 18*, 64-79
Cochran, W.G. (1977). *Sampling Techniques*. (3rd ed.). New York, USA: 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
Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology, 26*, 297-302
Fielding, A.H., & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation, 24*, 38-49
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
GCOS (2004). Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. In: World Meteorological Organization
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
Giglio, L., Randerson, J.T., & van der Werf, G.R. (2013). Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4). *Journal of Geophysical Research: Biogeosciences, 118*, 317-328
Hand, D.J. (1981). *Discrimination and Classification*. New York: John Wiley and Sons
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
Miller, R.G. (1966). *Simultaneous statistical inference*. London, United Kingdom: McGraw-Hill
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., Stehman, S.V., & Chuvieco, E. (2014a). 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. (2014b). Assessing the temporal stability of the accuracy of a time series of burned area products. *Remote Sensing, 6*, 2050-2068

Pereira, J.M., Mota, B., Calado, T., Oliva, P., & González-Alonso, F. (2013). ESA CCI ECV Fire Disturbance - Algorithm Theoretical Basis Document – Volume II – BA Algorithm Development. In: ESA Fire-CCI project

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., Quaife, 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

Roy, D., Jin, Y., Lewis, P., & Justice, C. (2005). Prototyping a global algorithm for systematic fire-affected area mapping using MODIS time series data. *Remote Sensing of Environment, 97*, 137-162

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., Arora, M., Kasetkasem, T., & Varshney, P. (2007). Estimation of Fuzzy Error Matrix Accuracy Measures Under Stratified Random Sampling. *Photogrammetric Engineering and Remote Sensing, 73*, 165-174

Tansey, K., Bradley, A., & Padilla, M. (2014). ESA CCI ECV Fire Disturbance - Algorithm Theoretical Basis Document - Volume III - BA Merging. In: ESA Fire-CCI project

Tansey, K., Brandley, A., Smets, B., van Best, C., & Lacaze, R. (2012). The Geoland2 BioPar burned area product. In E.G. Union (Ed.), *EGU General Assembly* (p. 4727). Vienna, Austria: Copernicus

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*

van der Werf, G.R., Randerson, J., T., Collatz, G.J., Giglio, L., Kasibhatla, P.S., Arellano, A.F., Olsen, S.C., & Kasischke, E.S. (2004). Continental scale-partitioning of fire emissions during the 1997 to 2001 El Niño/La Niña period. *Science, 303*, 73-76

Appendix: Estimated error matrices and accuracy measures describing accuracy for the globe considering product no-data as unburned, for all the selected TSAs (105).

Table A1: Estimated error matrices and reference burned area percent (BA%) for each product at the global level. Cells of the error matrices are represented by estimated percent area, 100* $\hat{P}_{ij}$. $\hat{P}_{11}$ represents the estimated proportion of area that is burned according to both the BA product and the reference classification, $\hat{P}_{12}$ is the estimated proportion of BA commission error, $\hat{P}_{21}$ is the estimated proportion of BA omission error, and $\hat{P}_{22}$ is the estimated proportion of area that is unburned according to both the BA product and the reference classification.

Product	$\hat{P}_{11}$	$\hat{P}_{12}$	$\hat{P}_{21}$	$\hat{P}_{22}$	BA%
MCD64	0.11	0.08	0.24	99.6	0.34
MCD45	0.09	0.08	0.25	99.6	0.33
Geoland2	0.03	0.09	0.30	99.6	0.33
MERGED_cci	0.07	0.45	0.26	99.2	0.33
MERIS_cci	0.08	0.14	0.25	99.5	0.33
VGT_cci	0.02	0.30	0.31	99.4	0.33

Table A2: Estimated global accuracy (expressed as %) of each product. Standard errors are shown in parentheses and superscript letters refer to products that have statistically significantly (α=0.003 per comparison) higher accuracy or lower bias than the designated product according to z tests.

Product	OA(%)	Ce(%)	Oe(%)	DC(%)	B(%)	relB(%)
a MCD64	99.7 (0.1)	43 (4)^{cdef}	69 (4)^{cdf}	40 (4)^{cdef}	-0.16 (0.05)	-46 (7)
b MCD45	99.7 (0.1)	46 (4)^{cdef}	74 (4)^{cf}	35 (4)^{cdf}	-0.17 (0.04)	-51 (7)

c Geoland2	99.6 (0.1)	76 (9)	92 (4)	12 (5)	-0.21 (0.07)	-65 (10)
d MERGED_cci	99.3 (0.2)	87 (5)[f]	79 (5)[cf]	16 (5)[f]	0.19 (0.17)	58 (53)
e MERIS_cci	99.6 (0.1)	64 (5)[df]	77 (5)[cf]	28 (5)[cdf]	-0.11 (0.04)	-35 (12)
f VGT_cci	99.4 (0.2)	93 (4)	94 (2)	6 (3)	-0.00 (0.17)	-1 (52)

502

Figure 1

Click here to download high resolution image

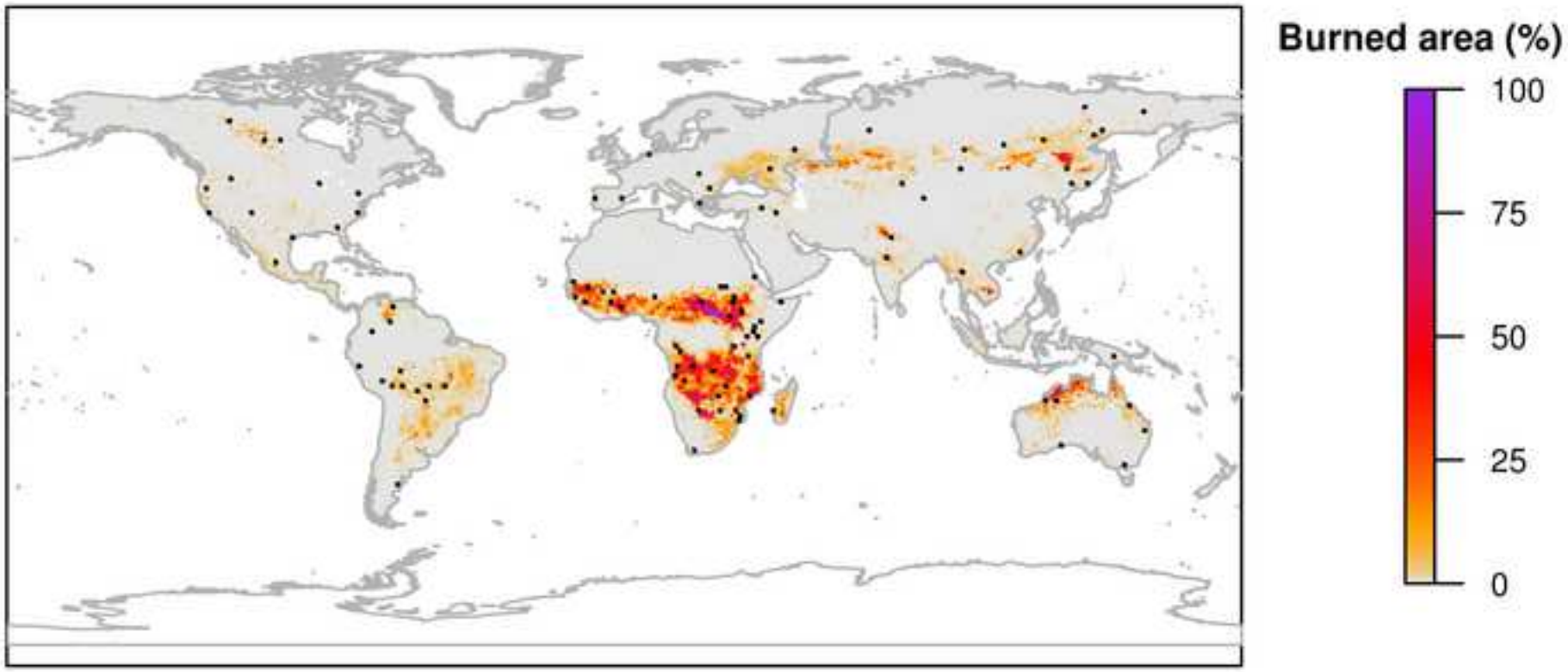


Figure 2

Click here to download high resolution image

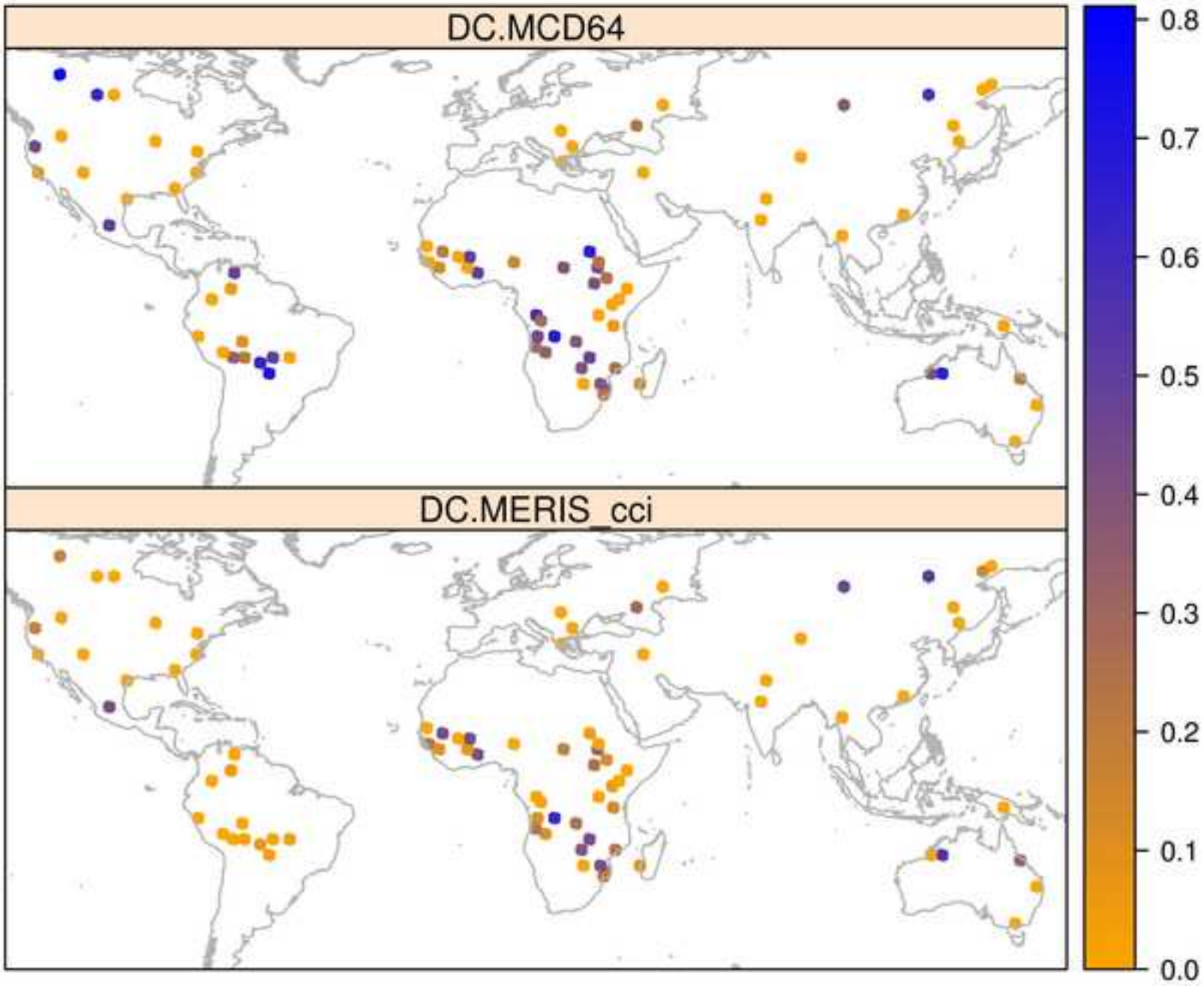


Figure 3

Click here to download high resolution image