Int J Appl Earth Obs Geoinformation 80 (2019) 295–311

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

# Global data and tools for local forest cover loss and REDD+ performance assessment: Accuracy, uncertainty, complementarity and impact

Astrid B. Bos[a,b,*], Veronique De Sy[a], Amy E. Duchelle[b], Martin Herold[a], Christopher Martius[b,c], Nandin-Erdene Tsendbazar[a]

[a] Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing, Droevendaalsesteeg 3, 6708 PB, Wageningen, the Netherlands
[b] Center for International Forestry Research, Jalan CIFOR, Situ Gede, Bogor Barat 16115, Indonesia
[c] Center for International Forestry Research (CIFOR) Germany gGmbH, Charles-de-Gaulle-Strasse 5, 53113, Bonn, Germany

ARTICLE INFO

*Keywords:*
Accuracy
Deforestation
Measurement, reporting and verification
Performance
REDD+
Uncertainty

ABSTRACT

Assessing the performance of efforts to reduce emissions from deforestation and forest degradation (REDD+) requires data on forest cover change. Innovations in remote sensing and forest monitoring provide ever-increasing levels of coverage, spatial and temporal detail, and accuracy. More global products and advanced open-source algorithms are becoming available. Still, these datasets and tools are not always consistent or complementary, and their suitability for local REDD+ performance assessments remains unclear. These assessments should, ideally, be free of any confounding factors, but performance estimates are affected by data uncertainties in unknown ways. Here, we analyse (1) differences in accuracy between datasets of forest cover change; (2) if and how combinations of datasets can increase accuracy; and we demonstrate (3) the effect of (not) doing accuracy assessments for REDD+ performance measurements.

Our study covers five local REDD+ initiatives in four countries across the tropics. We compared accuracies of a readily available global forest cover change dataset and a locally modifiable open-source break detection algorithm. We applied human interpretation validation tools using Landsat Time Series data and high-resolution optical imagery. Next, we assessed whether and how combining different datasets can increase accuracies using several combination strategies. Finally, we demonstrated the consequences of using the input datasets for REDD+ performance assessments with and without considering their accuracies and uncertainties.

Estimating the amount of deforestation using validation samples could substantially reduce uncertainty in REDD+ performance assessments. We found that the accuracies of the various data sources differ at site level, although on average neither one of the input products consistently excelled in accuracy. Using a combination of both products as stratification for area estimation and validated with a sample of high-resolution data seems promising. In these combined products, the expected trade-offs in accuracies across change classes (before, after, no change) and across accuracy types (user's and producer's accuracy) were negligible, so their use is advantageous over single-source datasets. More locally calibrated wall-to-wall products should be developed to make them more useful and applicable for REDD+ purposes. The direction and degree of REDD+ performance remained statistically uncertain, as CIs were overlapping in most cases for the deforestation estimates before and after the start of the REDD+ interventions. Given these uncertainties and inaccuracies and to increase the credibility of REDD+ it is advised to (1) be conservative in REDD+ accounting, and (2) not to rely on results from single currently available global data sources or tools without sample-based validation if results-based payments are intended to be made on this basis.

## 1. Introduction

Under the United Nations Convention on Climate Change (UNFCCC), reducing emissions from deforestation and forest degradation and enhancing forest carbon stocks (REDD+) has been initiated as an important climate change mitigation strategy. Hundreds of government and non-government led REDD+ programs and projects have emerged at the subnational and local level over the past decade

[*] Corresponding author at: Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing, Droevendaalsesteeg 3, 6708 PB, Wageningen, the Netherlands.
*E-mail address:* astrid.bos@wur.nl (A.B. Bos).

https://doi.org/10.1016/j.jag.2019.04.004
Received 13 December 2018; Received in revised form 7 April 2019; Accepted 9 April 2019

(Simonet et al., 2015). In order to track the performance of these initiatives, implementers must create or leverage measurement, reporting and verification (MRV) schemes for carbon stocks and carbon emissions. One approach to calculate carbon emissions is by multiplying the activity data in a given area by an emission factor (Verchot et al., 2012; IPCC, 2006). Activity data is the area of land changed from forest into another type of land use.

The estimation of activity data evolved rapidly through innovations in remote sensing and forest monitoring, with algorithms and datasets with ever increasing levels of coverage, spatial and temporal detail, and accuracy. However, these datasets do not necessarily agree with each other, and more transparency and better cooperation between the science and policy domain is required to measure –and realize– the mitigation potential of REDD+ activities (Grassi et al., 2017). Estimates can differ due to many factors, including misalignment of reference levels and time periods, forest and deforestation definitions used, and (remote sensing) data sources used for a map product (e.g. different satellite data) (Melo et al., 2018). Although the resulting differences in estimates are expectable and understandable, the ambiguity leaves room for political manoeuvring around the data (Wong et al., 2016) which threatens accountability. On the positive side, it is becoming more common practice to systematically report map product's accuracies and uncertainties (e.g. Olofsson et al., 2013, 2014; Stehman, 2014), increasing both transparency and product comparability. To this end, a reference classification is needed. Accuracy is defined as the degree to which the produced map agrees with this reference classification (Olofsson et al., 2013), which generally requires a sample-based validation. The uncertainty of the corresponding area estimates of, in this case, deforestation, is then expressed by the variance, standard error, or confidence intervals (CI) of these estimates. One could account for these uncertainties in the input data by being conservative about the subsequent REDD+ estimates, so as to prevent overestimation of the reduced emissions (Grassi et al., 2008).

Locally calibrated products are often favoured over global products, as this can considerably reduce the sample size for validation purposes (GFOI, 2016). Still, some widely used regional forest change datasets are found to be inaccurate by underestimating forest loss (Milodowski et al., 2017). Also, trade-offs exist between accuracy, local adjustability, and sample size needed on the one hand, and ease of use, processing time, knowledge and skills required on the other (Duchelle et al., 2015). While at the *national* level, in recent years the capacities of countries are increasing (Romijn et al., 2015), for *local* and *subnational* REDD+ initiatives it is often difficult and impractical to gain sufficient capacities and resources to perform proper area estimations. Here, the availability of open-source products provides an attractive opportunity. It remains understudied however, to what extent these readily available datasets and tools can contribute to challenges in the environmental domain and to REDD+ performance assessments in particular.

For local forest cover loss measurements, it is of vital importance to understand the differences in accuracies of forest cover loss maps derived from different products and tools. This supports the choice to use either more complex, time-consuming, but locally adaptable tools that provide the required high accuracies, or to opt for a readily available product with global coverage which might suffice in certain cases. In addition, accuracy assessment of combinations of products and tools can reveal their complementarities and show how uncertainties can be minimized while maximizing accuracies. In other words, in terms of increased accuracy and decreased uncertainty, a combined product may be better than the sum of its parts. An earlier study has focused on a comparison of available datasets in terms of in accuracy and uncertainty in one country (Melo et al., 2018), while others have studied the differences across several tropical countries (e.g. Turubanova et al., 2018). To the best of our knowledge, this is the first effort however, to compare different products at different (subnational) sites across the tropics, while exploring the potential and added value of combining those products.

Datasets used for REDD+ performance assessments should, ideally, be free of any confounding factors, but it is currently unclear how performance estimates are affected by data uncertainties. Hence, a systematic accuracy assessment is necessary to compare accuracies in various map products and to gain insight in the remaining uncertainty in deforestation area estimates. Furthermore, it remains understudied whether and how map products could complement each other and to what extent they are suitable for measuring the performance of REDD+. Therefore, the objectives of this study are to analyse if and how combinations of datasets can increase accuracy, and to understand how differences in accuracy between forest cover change datasets and its corresponding uncertainty influence REDD+ performance assessments. We defined the following research questions:

1) How do forest cover loss datasets differ in terms of accuracy?
2) What is the complementarity of these forest cover loss datasets in increasing accuracy?
3) How do map accuracy and area estimate uncertainty influence REDD+ performance assessment?

## 2. Methods & material

### 2.1. Study area

We use data from five local REDD+ initiatives located in four countries across the tropics (Table 1). These initiatives are part of the Global Comparative Study on REDD+ (CIFOR, 2017) and were selected to represent a wide range of intervention types ((dis)incentives and enabling measures), implementer types (government, non-governmental organization, private sector), and geographies across the

**Table 1**
Site characteristics.

| Site | (Approx.) size (ha) of area of interest (AOI) | Main ecozone(s) (source: FAO) | REDD+ start year | National forest definition[1] | |
|---|---|---|---|---|---|
| | | | | Tree cover (%) | MMU[2] (ha) |
| Peru | 1,100,000 | Tropical rainforest | 2009 | 30 | 0.09 |
| Tanzania | 200,000 | Tropical dry forest / tropical shrubland | 2010 | 10 | 0.50 |
| Vietnam | 800,000 | Tropical rainforest / Tropical moist deciduous forest | 2009 | 10 | 0.50 |
| Indonesia-A | 2,000,000 | Tropical rainforest | 2008 | 30 | 0.25 |
| Indonesia-B | 3,600,000 | Tropical rainforest | 2009 | 30 | 0.25 |

[1] Based on most recent submissions to UNFCCC (2019).
[2] Minimum mapping unit.

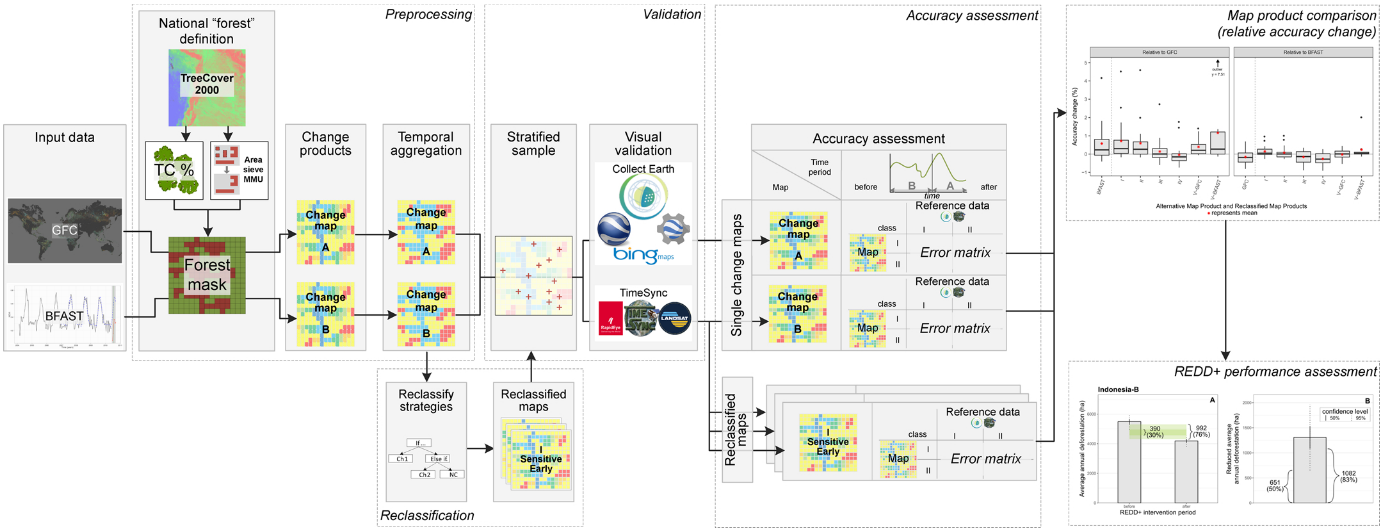


**Fig. 1.** Workflow and processing steps.

**Table 2**
Comparison of GFC and BFAST products (with information from Hansen et al., 2013; Verbesselt et al., 2012; Gross et al., 2017).

| | GFC | BFAST |
|---|---|---|
| Type | 2000 tree cover; loss; gain; and loss year raster products | Change detection algorithm |
| Sensor | Landsat ETM + | Depends on user input, here: Landsat ETM + |
| Spatial resolution | 30 m | Depends on user input, here: 30 m |
| Temporal resolution | Year | Julian day, limited by user input and cloud coverage |
| Spatial coverage | Global | Site based; 'case studies' |
| Algorithm | Bagged decision tree model | Additive season and trend model |
| Advantages | Global coverage, easy to use, end product freely available | Locally modifiable, open source |
| Disadvantages | Algorithm not flexible; not near-real time | Requires user's input data; requires expert knowledge; computationally intensive |
| Source | http://earthenginepartners.appspot.com/science-2013-global-forest | http://bfast.r-forge.r-project.org/ |
| Reference | Hansen et al. (2013) | Verbesselt et al. (2012) |

tropics. Furthermore, they vary in terms of size and environmental context, namely from dense primary rainforest to dry miombo woodlands (Sills et al., 2014). Data availability constraints affected the selection procedure, as the availability of both map products (section 2.3) was a prerequisite for this study.

### 2.2. Summary of workflow

The workflow and processing steps (Fig. 1) were repeated for each study site. We compared the accuracy of a tree cover change dataset, i.e. Global Forest Change (GFC), and a map developed using an open-source algorithm to detect forest cover change, i.e. Breaks For Additive Seasonal and Trend (BFAST). For each study site and based on national forest definitions, we used the same forest mask using tree cover (TC) percentage and an area sieve using the minimum mapping unit (MMU) (Table 1). We thus compared differences in change detection between the two input products, rather than differences in forest definitions applied. We considered three classes: *before*, *after* and *no change*. The transition between *before* and *after* is defined by the start year of each studied initiative. We combined the two products using different reclassification strategies, which led to a set of new combined change map products. We applied a stratified random sample on the change map and validated the original products and reclassified products using a set of visual tools. Accuracies were calculated using these validation samples, as well as the differences in accuracies relative to the two input map products. The uncertainty in the area estimates was expressed using the 95% and 50% CIs of those estimates. We compared the map estimates and reference-based area estimates. Finally, we assessed the influence of uncertainty in the area estimates and their trends on REDD+ performance measurements. All analytical steps are discussed in more detail below.

### 2.3. Input data

For the first map product, we used Global Forest Change (GFC) data (version 1.3), a Landsat-based time-series dataset of tree cover density in 2000 and annual tree cover loss for 2001–2015 (Hansen et al., 2013). The GFC product provides yearly forest cover loss data with global coverage. Together with baseline data on forest cover in 2000, users can relatively easily examine deforestation patterns using their own forest

**Table 3**
Temporal alignment of change products per study site.

| Site | Time frame GFC | Time frame BFAST | REDD+ start year | Aligned *before* period | Aligned *after* period |
|---|---|---|---|---|---|
| Peru | 2001-2015 | 1999-2014 | 2009 | 2001-2008 | 2009-2014 |
| Tanzania | 2001-2015 | 2005-2015 | 2010 | 2005-2009 | 2010-2015 |
| Vietnam | 2001-2015 | 2005-2014 | 2009 | 2005-2008 | 2009-2014 |
| Indonesia-A | 2001-2015 | 2001-2014 | 2008 | 2001-2007 | 2008-2014 |
| Indonesia-B | 2001-2015 | 2001-2015 | 2009 | 2001-2008 | 2009-2015 |

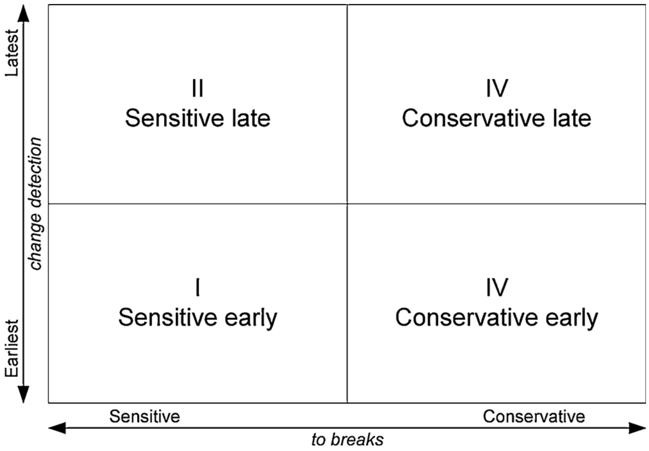


**Fig. 2.** Rationale behind reclassification strategies.

definitions. Data analysis using the Global Forest Watch tools does not require expert GIS knowledge.

The other product is based on the Breaks For Additive Seasonal and Trend (BFAST) algorithm (Verbesselt et al., 2010, 2012; DeVries et al., 2015), which requires a time series of local input data (here, NDVI and NDMI based on Landsat satellite data). With this adaptable open-source deforestation detection algorithm, users can analyse deforestation patterns in their own time series data in, for example, a cloud processing environment. It is usually applied to smaller areas, as processing time increases with longer time series and larger area spans. Some degree of remote sensing knowledge and coding skills are necessary to apply the algorithm on the time series. The algorithm is highly flexible and can be adapted to the local (environmental) context and user needs. The user can calibrate the model by adjusting the parameters to the local context, resulting in change rasters with interannual precision. Both products allow the user to create forest cover change products with a temporal resolution of one year or shorter, and a spatial resolution of 30 m. The main differences between the two products regard their flexibility, coverage, and ease-of-use (Table 2).

### 2.4. Pre-processing

We aligned our forest definitions with the corresponding countries' definitions[1] . These generally consist of a tree cover or crown percentage at the baseline year and a minimal mapping unit (MMU) (Table 1). GFC's tree cover density layer for the year 2000 (TC2000) allowed us to create forest masks based on the nationally defined tree cover percentage thresholds. Next, we applied area sieves following the countries' defined MMU and applied these forest masks to both input products. We defined deforestation as a change from forested land (using the forest mask) to land that has been clear cut (i.e. bare soil)[2] .

In addition to aligning forest and deforestation definitions, we needed to temporally align the data for the products to represent the same time periods (Table 3). We then aggregated the change products into three classes, representing (1) the period *before* the REDD+ interventions started, (2) the period *after* the interventions started, and (3) *no change* (i.e. stable forest). All other pixels, (i.e. non-forest; forest cover change in other years etc.) were excluded from further analyses.

### 2.5. Reclassification of change products

Since these datasets generally have their own strengths and weaknesses (Table 2), we assessed whether joint products can lead to an accuracy increase. Therefore, we combined the two products at pixel level using five different reclassification strategies. The first four strategies are defined by differences in sensitivity to change and in timing of change detection (Fig. 2), based on the following decision rules:

- *I Sensitive early* – Adopt value of change product that detects a disturbance the earliest, regardless of the other change product's detection;
- *II Sensitive late* – Adopt value of change product that detects a disturbance the latest, regardless of the other change product's detection;
- *III Conservative early* – If any of the change products classifies the pixel as *no change*, then the decision for the reclassified product is *no change*. If both products detect change, trust the earliest detection;
- *IV Conservative late* - If any of the change products classifies the pixel as *no change*, then the decision for the reclassified product is no change. If both products detect change, trust the latest detection.

A fifth strategy was added to represent a case in which the timing of change detection is irrelevant. Here, the two individual products were aggregated into two binary *change-no change* rasters, disregarding the year or corresponding period of change detection. Details are visualized in Appendix A.

Table 4 shows the reclassification strata for each strategy, which formed the input for the stratified sampling (see next section). For each site, the five reclassification strategies resulted in six extra change maps, that is, four combined and two 'timeless' raster datasets, which were added to the accuracy assessment for comparison with the original GFC and BFAST products.

### 2.6. Validation

Sample size is important when designing validation schemes for comparative purposes (Foody, 2009). Although our individual

[1] Following the submissions to the UNFCCC's REDD+ platform (UNFCCC, 2019).

[2] Sometimes land use change from (natural) forest to forest plantation is considered degradation or even enhancement of carbon stock (e.g. in Vietnam's REDD+ FRL submission to the UNFCCC, 2016), but here it is considered deforestation, since at –at least- one point in time the forest was cleared which leads to a reflectance of bare soil.

**Table 4**
Strata and classification values of different reclassification strategies.

| | | | Combination strategies | | | | Timeless strategies | |
|---|---|---|---|---|---|---|---|---|
| | | | I | II | III | IV | V | |
| GFC | BFAST | Validation stratum | sensitive – early | sensitive – late | conservative – early | conservative – late | timeless-GFC | timeless-BFAST |
| before | before | 1 | before | before | before | before | change | change |
| after | before | 3 | before | after | before | after | change | change |
| no change | before | 4 | before | before | no change | no change | no change | change |
| before | after | 3 | before | after | before | after | change | change |
| after | after | 2 | after | after | after | after | change | change |
| no change | after | 5 | after | after | no change | no change | no change | change |
| before | no change | 4 | before | before | no change | no change | change | no change |
| after | no change | 5 | after | after | no change | no change | change | no change |
| no change | no change | 6 | no change | no change | no change | no change | no change | no change |

aggregated change raster datasets consisted of three classes (*change before*, *change after* and *no change*), for simplification in the sampling design we considered them as having a binomial distribution (either *change* or *no change*) and used an alpha of 0.10, planned proportion estimate of 0.5 (i.e. conservative), and 0.05 margin of error leading to a sample size of 270 pixels per site (Foody, 2009; Cochran, 1977).

We overlaid the two input change products with each three classes, resulting in nine possible combination values. These nine classes were aggregated into six strata (Table 4). At each site, the 270 pixels were randomly selected across the strata, which led to 45 sample pixels per stratum (Fig. 3).

A validation survey was developed using Open Foris Collect (Open Foris, 2019). The survey and samples were loaded into Google Earth via CollectEarth and simultaneously visualised in R using the TimeSync package (Cohen et al., 2010). Each sample was visually checked through multiple available historical images within Google Earth (if any), the most recent Bing Maps image, the most recent image via Google Earth Engine, and false colour yearly composites of Landsat data within Google Earth Engine. Within R, a time series of RGB and false colour (NIR, SWIR1, red) snapshots were created with TimeSync. Together this allowed us to determine (1) whether there was any disturbance and, (2) if so, to find the timeliest disturbance date. In case of multiple disturbances within the time series, the first disturbance was recorded.

### 2.7. Accuracy assessment

After completing the validation survey, the visual judgements from the validation survey were compared with the findings from the GFC, BFAST and reclassified products. A map pixel was considered correct if both the status (*change* or *no change*) and time period (*before* or *after*) matched the visual judgement. Accuracies of the map products and the class area proportions were estimated while taking into account the inclusion probability of the samples per site. Since the sampling stratification was a combination of GFC and BFAST results, we followed the approaches detailed in Stehman (2014) which addresses estimating map accuracies and class areas when the sampling strata are different from the map classes. CIs of the estimation also followed the same method (Stehman, 2014; Cochran, 1977). For the remainder of this article, with ‘map-based area estimates’ we refer to area estimations directly calculated from the maps, whereas ‘reference-based area estimates’ refers to the areas as derived from the class area proportions coming from the sample-based validation using reference data.

Next, the differences in overall, producer’s (inversely linked to errors of omission) and user’s (inversely linked to errors of commission) accuracies were assessed by calculating the *relative accuracy changes*, which give insight in which reclassification strategy provides the largest increase in accuracy compared to the original input products. Relative accuracy change was calculated as follows:

$$\boldsymbol{RA}(\boldsymbol{x}) = \frac{\boldsymbol{A_x} - \boldsymbol{A_y}}{\boldsymbol{A_y}} \tag{1}$$

Where $x$ is the alternative map product, $y$ is the original map product (either GFC or BFAST), and $A$ is the corresponding accuracy (overall, producer’s or user’s accuracy).

### 2.8. Performance assessment

In this study, we simplify REDD+ performance by referring to the direction in deforestation trend over time, hence good REDD+ performance corresponds to reduced average annual deforestation. We compared the trends in average annual deforestation from before and after the start of the REDD+ intervention (Bos et al., 2017). The impact of ignoring data accuracy in REDD+ performance assessments was assessed by comparing the average annual deforestation per period for the map estimates and reference-based area estimates.

Trends and uncertainties were assessed in two ways. First, they were visually assessed by focusing on the overlap of the CIs of the deforestation estimates in the *before* and *after* period. Presence of such overlap would mean that direction and magnitude of REDD+ performance remains uncertain. Absence of such overlap would reveal the direction of deforestation trend and its magnitude with more certainty. In addition to the commonly used 95% CI, we applied a 50% CI. This means one accepts a 25% probability of overestimating the ‘true’ REDD+ value in the monitoring period, which is similar to the adjustment procedure under Article 5.2 of the Kyoto Protocol (UNFCCC, 2006, cited in Grassi et al., 2008). Second, the trend uncertainty was calculated using the (joint) variances and CI of the trend itself (GOFC-GOLD, 2016).

The conservativeness principle (Grassi et al., 2008) was applied to a case with a decreasing trend, to examine the influence of different conservativeness standards on the final REDD+ estimate.

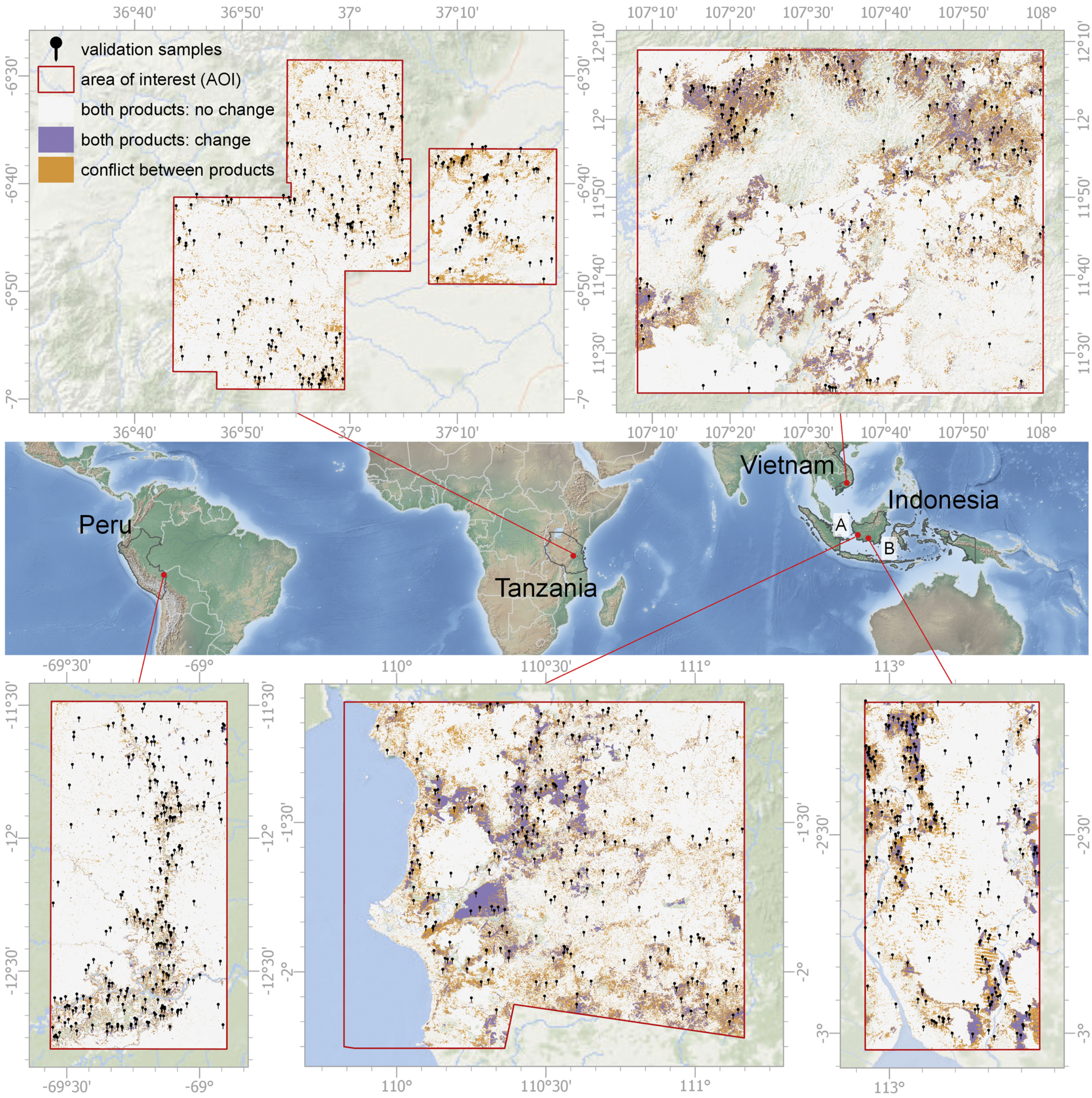


**Fig. 3.** Study sites with validation samples and areas of agreement and conflict between the two input map products. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

## 3. Results

### 3.1. Annual deforestation rates

Fig. 4 shows an overview of the annual deforestation rates for both GFC and BFAST input products at each site before the accuracy was assessed and thus before the area estimates of deforestation using the reference data were calculated. Both products show overall higher annual deforestation rates in the southeast Asian sites compared to the sites in Peru and Tanzania. The deforestation trends appear similar when comparing the two products at all sites. However, deforestation estimates in individual years differed considerably, especially so in

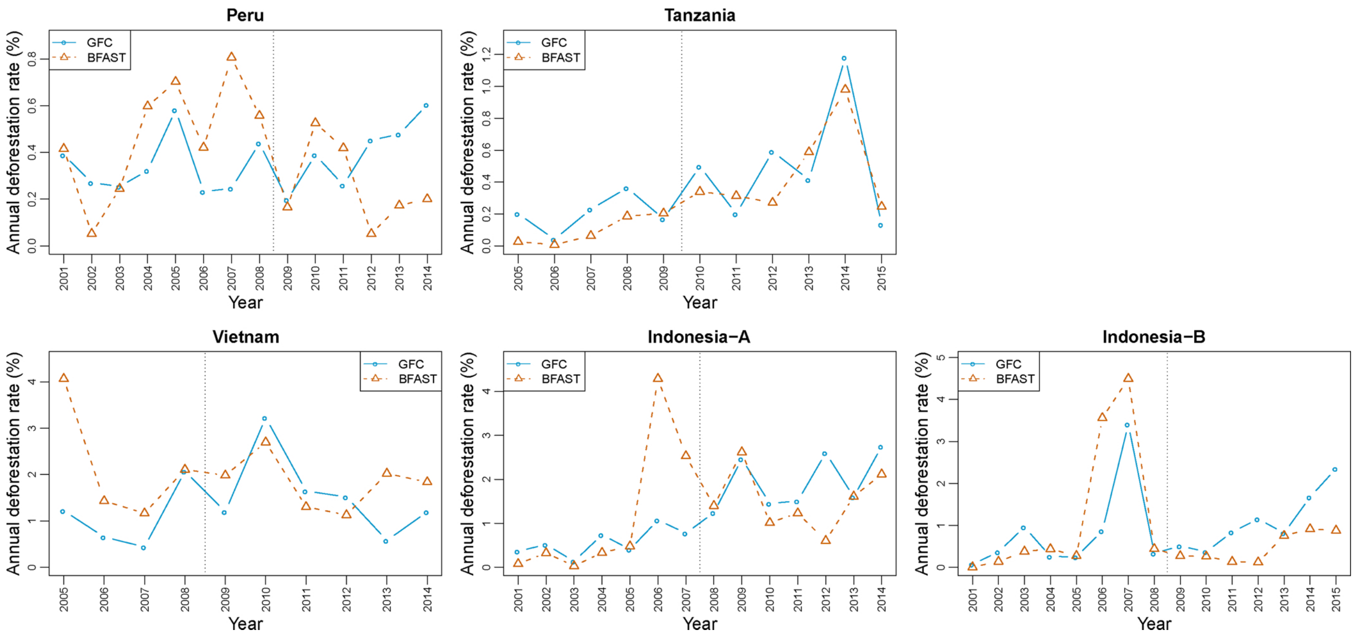


**Fig. 4.** Site-based comparison of annual deforestation rates. Rates represent the deforestation detected by the input products as percentage of forest cover in 2000. Note that the x and y-scales differ per site. The vertical dotted line represents the start year of the REDD+ intervention(s) in the corresponding site and thus the transition from the *before* to *after* period.

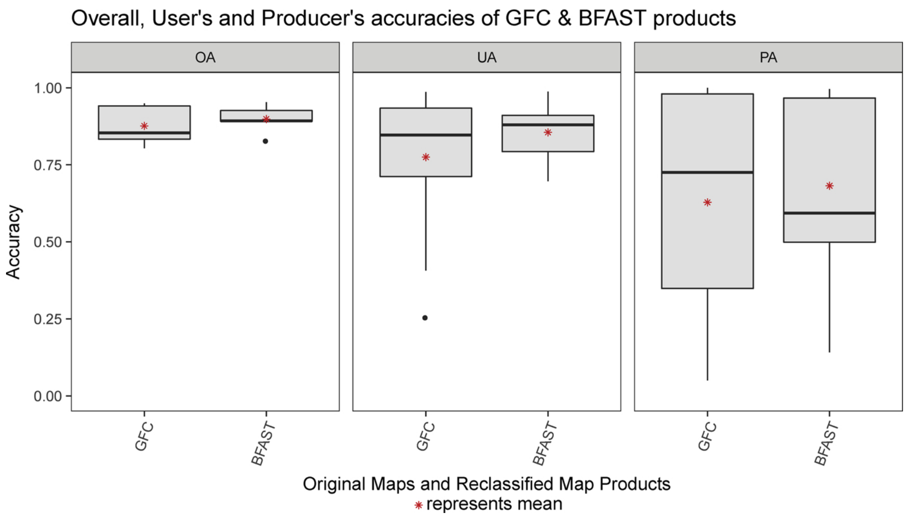


**Fig. 5.** Overall (OA), user’s (UA) and producer’s (PA) accuracies of GFC and BFAST products. All classes (i.e. *change before*, *change after*, and *no change*) are included. Upper and lower extremes of whiskers represent Q3 + 1.5* interquartile range (IQR) and Q1–1.5*IQR respectively, where IQR = Q3 − Q1.

Vietnam (2005) and Indonesia (2006 and 2007), which might indicate differences in timeliness of deforestation detection. In terms of REDD+ performance, these results reveal some ambiguity of the deforestation trends. In Peru, the GFC showed slightly increasing deforestation while according to BFAST deforestation was generally going down since the start of the REDD+ initiative. The site in Tanzania showed no clear performance while the steep drop in deforestation in site Indonesia-B after 2007 might indicate positive REDD+ performance.

### 3.2. *Accuracy*

#### 3.2.1. *Overall, user’s and producer’s accuracy*

For all original and reclassified map products, the error matrices were calculated based on the comparison between the map class (*change before*, *change after*, *no change*) and the visually assigned class using the reference data. Fig. 5 shows the overall (OA), user’s (UA) and producer’s (PA) accuracies stemming from these error matrices.

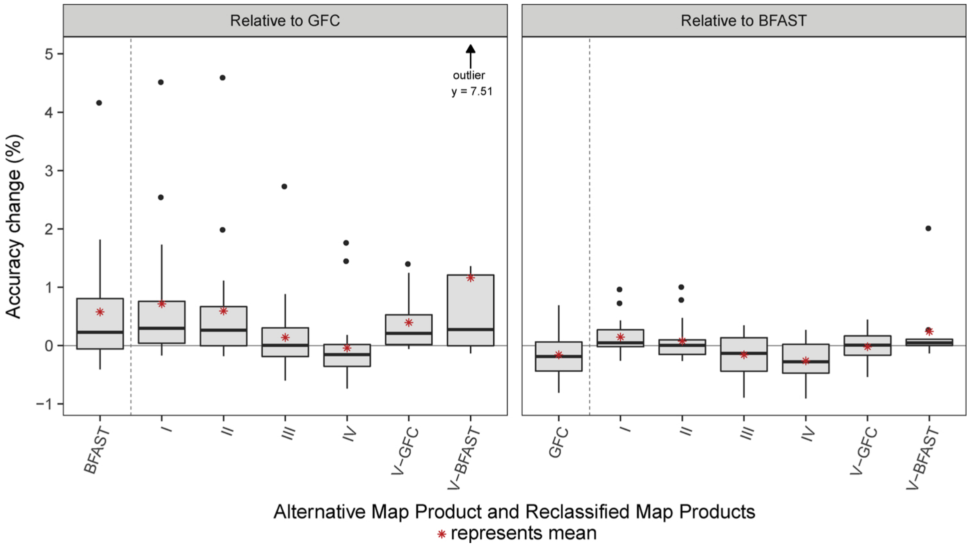


**Fig. 6.** Relative change of the accuracies per alternative product. The figure includes the accuracies (only PA and UA) of the before and after change classes of all sites. Upper and lower extremes of whiskers represent Q3 + 1.5*IQR and Q1–1.5*IQR respectively, where IQR = Q3 − Q1.

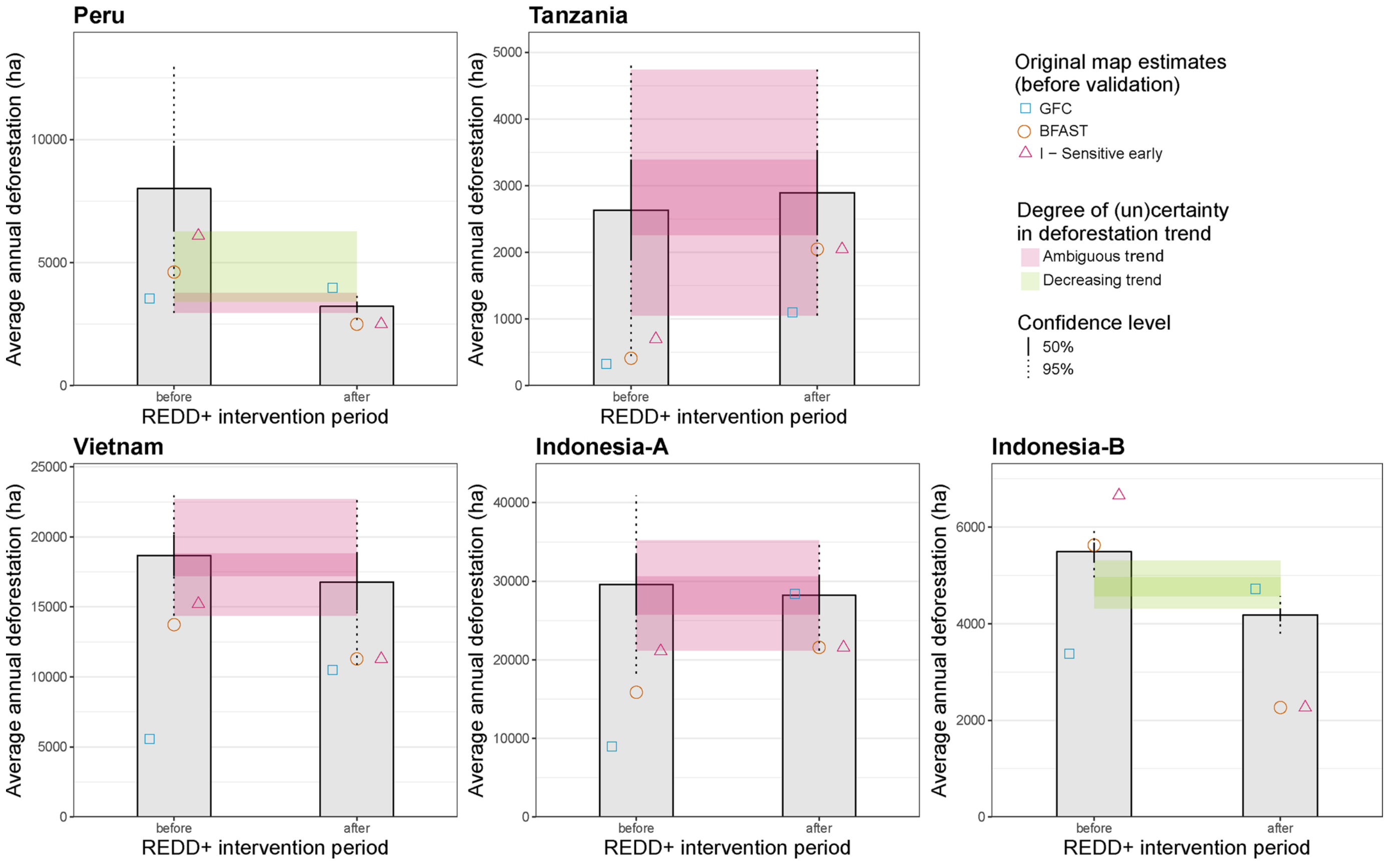


**Fig. 7.** Influence of accuracy assessment and area estimates' uncertainty on REDD+ performance measurements. The grey bars represent the average annual deforested areas (reference-based area estimates), with 95%CIs. We corrected the CIs for differences in the number of years between the *before* and *after* period, assuming variances to be equally distributed in time. The selection of best performing reclassified product is based on the highest relative accuracy change, excluding the two V-timeless reclassified products, leading to I-sensitive early for all sites. The pink shaded areas represent the remaining degree of uncertainty, in which the direction of the deforestation trend remains ambiguous after considering the accuracy assessment. There is no overlap in the CIs of Peru (50%CI) and Indonesia-B (both 50%CI and 95%CI), hence the absence of a pink shaded area. The green shaded areas in those sites represent the downwards trend in deforestation, without overlap of CIs (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

**Table 5**
Direction and degree of deforestation trend for the GFC, BFAST and reclassification strategy I map estimates, and for the area estimates using reference data. Trend uncertainty is indicated for the reference-based area estimates.

| | GFC[1] | BFAST[1] | I-Sensitive Early[1] | Area estimates[2] | | |
|---|---|---|---|---|---|---|
| | change (%)[3] | change (%)[3] | change (%)[3] | change (%)[3] | Trend uncertainty with 95%CI (%.)[4] | Trend uncertainty with 50%CI (%.)[4] |
| Peru | 12 | −46 | −59 | −60* | 63 | 22 |
| Tanzania | 238 | 397 | 192 | 10 | 109 | 37 |
| Vietnam | 89 | −18 | −26 | −10 | 39 | 13 |
| Indonesia-A | 217 | 36 | 2 | −5 | 45 | 16 |
| Indonesia-B | 40 | −60 | −66 | −24** | 12 | 4 |

1 Original map estimates before accuracy assessment.
2 Area estimates based on the reference data, see also. Fig. 7.
3 Degree of change (%) when comparing average deforestation in *after* period with average deforestation in *before* period. A negative number signifies a decrease in (average annual) deforestation over time.
4 Uncertainty (U) of the trend in percent points is calculated as follows: U = CI(ha)/def_bef(ha). Where.
CI(ha) = sqrt(((var_bef/n_befˆ2) + (var_aft/n_aftˆ2))*TotalAreaˆ2)*z.
var_bef and var_aft are the variance of the area proportion of the classes before and after respectively.
n_bef and n_aft are the number of years in the before and after period respectively.
z is 1.96 and 0.67 for the 95%CI and 50% respectively.
def_bef(ha) is the estimated deforestation in the before period in hectares.
* no overlap between 50%CI of the before and after estimates (here: decreasing deforestation trends).
** no overlap between 50%CIs and 95%CIs of the before and after estimates (here: decreasing deforestation trends).

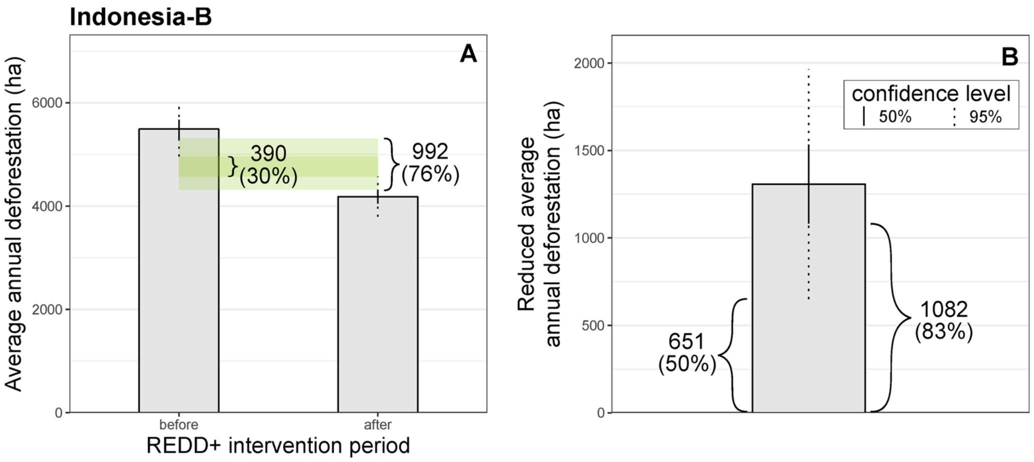


**Fig. 8.** Conservativeness principle applied to calculate the REDD + estimates for Indonesia-B. With approach A (left) one prevents overestimation of the reference estimates (before period) and underestimation of the assessment period (after period). Estimates in approach B (right) are derived from the uncertainty of the trend. Numbers next to curly brackets show the conservative REDD + estimate (activity data only) in ha assessed at the 95%CI and 50%CI, and as percentages of the trend from the reference-based area estimates (grey bar in B) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

While the OAs for all products were high, this result primarily stems from correctly detected *no change* areas (stable forest cover) which spans the majority of the areas studied. The UAs and PAs, and their corresponding errors of commission and omission respectively, were more informative for change related classes. In general, the lower PAs indicate that all products underestimate deforestation. When comparing the two input products, BFAST shows on average slightly higher accuracies (OA, UA and PA) compared to the GFC product. We found a general tendency of lower variation in accuracies of BFAST as compared to GFC across the sites.

However, whether GFC or BFAST performed better in terms of accuracy differs per study site (Appendix B). In the Peruvian site, the GFC and BFAST accuracies were quite similar, although there were some notable differences in PA. The Tanzanian site was characterized by low accuracies in general, but BFAST seemed to perform better at distinguishing real deforestation impacts from seasonal effects, hence the difference in UA between the two products. In both Indonesian sites the PA of BFAST in the after class was lower compared to GFC, indicating that the most recent changes are not well detected by BFAST. In the Vietnamese site, this was the opposite, as the PA of BFAST outperformed GFC in the after class.

### 3.2.2. Relative accuracy change

Comparing accuracies of the original map products and the reclassified map products based on the five strategies (section 2.5), in four out of five sites[3] combining input maps following a *sensitive-early* strategy led to significantly higher accuracies[4] compared to the original map products alone (Fig. 6, Appendix C and Appendix D). Still, in the Tanzanian site, none of the reclassification strategies led to higher accuracies compared to (one of the) individual datasets, due to the poor performance of the GFC product in this study area.

[3] With the Tanzanian site being the exception.
[4] Increases in OA and PA in the change classes, with non to only slight (insignificant) decreases in UA, significance level 0.95. (Appendix C).

### 3.3. REDD+ performance assessment

#### 3.3.1. Revealing the deforestation trend

To assess the influence of the map products' accuracies and area estimate uncertainties on REDD+ performance assessments, we visualize the average annual deforestation (in ha) in the *before* and *after* class (Fig. 7). The results show that both the magnitude and trend of deforestation delineated from the map products differed greatly from reference-based area estimates. In Peru, Tanzania and Vietnam, the map-based deforestation estimates of the reclassified map product are closer to the reference-based estimates than those of the two original products. In addition, in four out of five cases[5] the reclassified product reveals the same deforestation trend as the reference-based area estimates, although the magnitude of change differed (Table 5). This reflects the added value of using a combined product over a single product, although accuracy assessment thus remains necessary. As Table 5 shows, in three out of five sites the direction of the deforestation trend according to the best reclassified product was different from at least one of the individual products, which would have had major implications if results-based payments would be based on a single product alone and disregarding the product's map accuracies and estimate uncertainties.

The majority of the map-based estimates (both the two input products and reclassified product) fell outside the 95%CI of the reference-based area estimates of both change classes, which affirms the importance of doing a (sample-based) validation of the map products. At the Indonesia-B site, the accuracy assessment elucidated the direction of performance considerably, as the 95%CIs around the reference-based area estimates are relatively small. Here, the average annual deforestation decreased from the *before* to the *after* class, while the corresponding CIs did not overlap, indicating a clear downwards trend in deforestation (green shaded area in Fig. 7, Table 5). At the site in Peru, both CIs in the *before* period are relatively large, but at a 50%CI a clear downwards trend in deforestation was found, as the CIs did not overlap. In all other sites, uncertainty in the direction of performance remained, since the CIs of the *before* and *after* period overlapped, as illustrated by the pink shaded area in Fig. 7.

#### 3.3.2. Uncertainty of the deforestation trend

In addition to the visual assessment of uncertainty of the trend, we quantified the uncertainty of the trend's magnitude. Therefore, we estimated the trend uncertainty (two rightmost columns of Table 5), which is based on the joint variance of the two monitoring periods, and is expressed in percent points (GOFC-GOLD, 2016).

As an example, in Vietnam the reference-based area estimates revealed an average annual *decrease* in deforestation of 10% with a trend uncertainty of ± 13% points at the 50%CI. Thus, at this confidence level an actual *increase* in deforestation of 3% is one of the possibilities. In Indonesia-B, the absence of overlapping confidence levels revealed a downwards trend. According to our area estimates this average annual decrease is 24% with a trend uncertainty of ± 12 and ± 4% points at a 95%CI and 50%CI respectively.

#### 3.3.3. Applying the conservativeness principle to the REDD+ estimate – a case study

As illustrated above, deforestation estimates are subject to uncertainty, which is why one should be conservative when accounting for REDD+ in order to increase its credibility despite those uncertainties (Grassi et al., 2008). In other words, one should take into account the data uncertainties to prevent overestimation of the reduced emissions. Grassi et al. (2008) present four approaches to account for data uncertainty using the conservativeness principle, of which we apply two (A2 and B1 in Grassi et al., 2008, here referred to as A and B respectively) to our deforestation estimates of Indonesia-B. As Fig. 8 shows, both the approach and confidence level chosen have a great impact on the REDD+ estimate, with conservative estimates of reduced annual deforestation ranging from 390 to 1082ha, or 7 to 20% respectively.

## 4. Discussion

Most likely any deforestation map contains classification errors (Olofsson et al., 2013), and deforestation area estimates from these maps would thus differ from reality. We showed that a systematic accuracy assessment is critically important to value the usefulness of wall-to-wall forest cover change datasets for local REDD+ performance measurements. Distinguishing between overall, user's and producer's accuracy allowed comparison of different maps and helped to understand to what extent a map is likely to over- or underestimate real deforestation. The subsequent analysis of the variances and CIs showed to what extent the deforestation estimates remained uncertain. Furthermore, in this multi-site analysis, we assessed if, how and where a combination of forest cover change datasets can help to increase the accuracy and reduce the uncertainty of deforestation estimates for measuring the performance of local REDD+ initiatives.

We found high overall accuracies but striking differences in user's and producer's accuracies and area estimates, which is in line with findings from Melo et al (2018) in Guinea-Bissau. In our multi-site study, however, large regional differences appeared in the degree of discrepancy between the map products, with notable differences in the producer's accuracies particularly. Several recent changes were missed by BFAST leading to a lower PA in the *after* period, while BFAST's PA outperformed the GFC product in the first years of the monitoring period. Combining forest cover change datasets using a *sensitive-early* strategy generally improved accuracies and reduced uncertainties despite expected trade-offs between different types of accuracies. That is, as expected, in three of the sitesa *sensitive-early* strategy led to slightly lower user's accuracies in the change classes due to a small increase in commission errors, but OAs and PAs increased more than the UAs deteriorated. Still, only in cases where the individual datasets showed reasonable to good accuracies, combining datasets led to a map product that was more accurate than the individual datasets, as low accuracies in one dataset could not be compensated by high accuracies in the other.

We found differences in timeliness of deforestation detection between GFC and BFAST, although these differences were not unidirectional across all sites. As stated in section 3.2.1, in the Indonesian sites, the lower PA of BFAST indicates omission errors in the *after* class, while in the Vietnamese site, BFAST appears to detect recent changes better than GFC does. Both GFC and BFAST appeared to have issues with a timely detection of deforestation due to mining, leading to errors of omission, while the visual validation with false-colour images showed easily detectable changes. More research is needed to verify if there is a correlation between the time series bands and corresponding vegetation and moisture indices, and their fitness to detect mining.

[5] Indonesia-A being the exception, here the area estimates showed a slight decrease, while the I-sensitive early product showed a slight increase.

With our stratified sampling design (section 2.4) there was less risk of overlooking missed deforested pixels (i.e. missed omission errors), as conflicting pixels (in which change is detected in one, but not in the other product) were included in the sample as a separate stratum. On the downside, this might have led to an overestimation of omission errors due to the large area weight of stable forest classes in stratified sampling. Although our error matrices accounted for disproportional sampling of conflicting pixels, it is still likely that the producer's accuracies of both products were negatively influenced by this sampling design. At the same time, due to our sampling design we may have missed some omission errors in the non-forest class, i.e. pixels that were (erroneously) not included in the initial forest mask but in fact deforested. We focus on the (in)correct classification of change or no-change *within* the (initial) forest however, rather than the initial classification of forest or non-forest. We thus compared the change products in itself, and not differences in (or the accuracy of) forest masks. As the reference-based area estimates are only based on the reference samples due to the applied method (Stehman, 2014), the sampling design has a great influence on the results. Increasing the sample size further would reduce the uncertainty in the area estimates.

We focused on the uncertainty in performance assessments as caused by the underlying forest cover change dataset(s). Yet, uncertainty may come from more sources, including the precision and influence of the REDD+ initiative start year. We aggregated the (sub) annual deforestation detections into three classes: change *before* REDD +; change *after* REDD+ started; and *no change*. This rather sudden, and mainly theoretical, transition from the *before* to *after* class may have influenced our accuracy estimates. In practice, many local REDD+ initiatives are continuations of earlier integrated conservation and development projects, so interventions towards protecting local forests may predate the official start dates (Sunderlin et al., 2015). Since transitions in forests and forest use are often gradual processes too, this complicates performance assessments even further. Longer time series may be needed to clearly show the impact. Finally, all accuracies calculated are relative to the reference dataset, which in this case was created through the visually validated samples. Errors in the classification through visual validation were limited by using multiple time series data sources (e.g. RapidEye, Landsat TM) and multiple tools (i.e. TimeSync and CollectEarth).

It is important to note that for each site, a right-angled AOI was defined using the initiative's boundaries and a buffer (Fig. 3). Therefore, the AOIs included more than the 'pure' REDD+ intervention areas. Our objective was to explore the potential of combining activity datasets for accuracy improvement, and to demonstrate the implications of ignoring data uncertainties for performance measurements, rather than to calculate (change in) deforestation and corresponding carbon emissions for individual sites or to assess actual performance of specific initiatives. The results presented in section 3.3 should therefore not be used to assess the performance of these REDD+ initiatives as such.

## 5. Conclusion

We analysed the differences in accuracy and uncertainty between two forest cover change datasets for five sites and studied if and how combinations of datasets can increase accuracies and reduce uncertainties in the context of local REDD+ performance assessments. We demonstrated the use and usefulness of these global products to assess forest cover loss at the local level.

*How do forest cover loss datasets differ in terms of accuracy?*

We found that accuracies differ at the site level, although on average neither GFC nor BFAST excelled in accuracy. In the sites in Peru, Tanzania and Vietnam, BFAST performed better, while in the Indonesian sites, GFC achieved higher accuracies. Both GFC and BFAST underestimated deforestation, as reflected by the lower producer's accuracies and corresponding higher errors of omission.

*What is the complementarity of these forest cover loss datasets in increasing accuracy?*

Knowing the strengths and weaknesses of the individual products, we assessed their complementarity by overlaying the two products using different reclassification strategies. The strategy that led to the *highest* accuracy increases and uncertainty decreases differed per site. In four out of five cases, a *sensitive-early* strategy led to higher accuracies compared to the individual products. Only when both products' individual accuracies were already reasonable to good, a reclassification strategy resulted in higher accuracies. Products with low accuracies could not be ameliorated by any of our reclassification strategies.

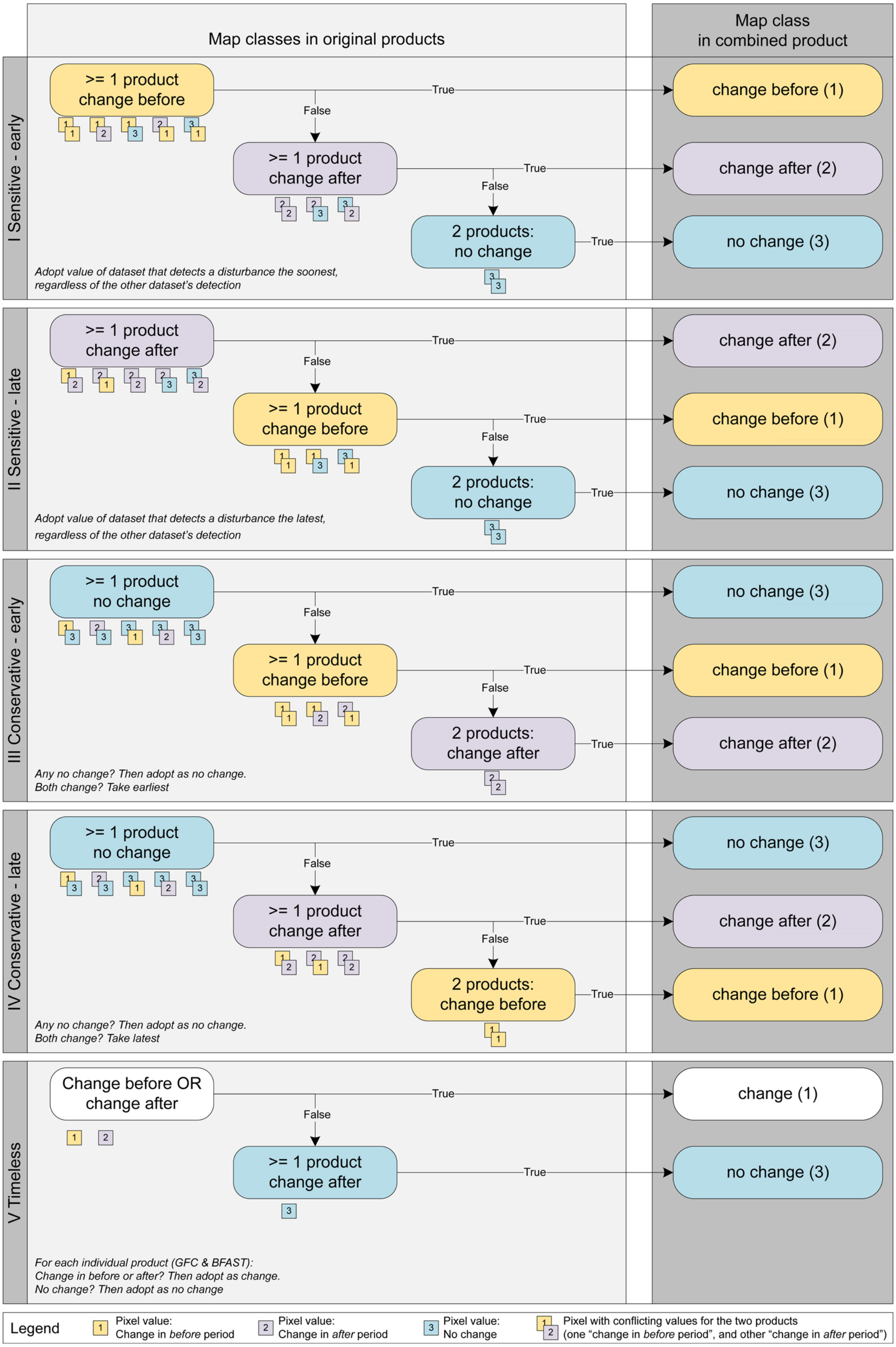
*How do map accuracy and uncertainty influence REDD+ performance assessment?*

We show the influence of input data accuracies and remaining uncertainties in annual deforestation estimates on REDD+ performance assessment and demonstrate the importance of accuracy assessment. As the overlap in CIs indicated, in three out of five sites some degree of uncertainty in the deforestation trend remained, even after accuracy assessment. In one site, the accuracy assessment revealed a clear downwards trend in deforestation. In one other site, the (absence of a) clear downwards trend was dependent on the confidence level chosen. In three sites, the annual deforestation estimates of the reclassified product were closer to the reference-based estimates when compared to the estimates of GFC and BFAST. Still, these map-based estimates were mostly outside the 95%CI of the reference-based estimates, thus affirming the persistent need for validation. But even reference-based estimates are subject to uncertainty, thus leading to a need for conservative accounting of corresponding REDD+ estimates. The growing availability of global, readily available datasets and tools is of vital importance as local implementers' monitoring capacities are often limited. Our comparative study shows that consideration of and transparency about accuracies, (un)certainties and corresponding (dis)abilities of datasets and tools, is of key importance if results-based payments are to be based upon performance measurements. Being conservative in REDD+ accounting could help address these uncertainties and thus increase the credibility of the REDD+ estimates. To get insights into, and ultimately reduce, uncertainty, we showed that the value of sample-based accuracy assessments cannot be overstated.

## Acknowledgements

The authors thank Tom Bewernick, Mathieu Decuyper, Ben DeVries and Erik van Schaik for their help with data processing; Mark Jorritsma for his assistance with the accuracy assessment; Sytze de Bruin and Valerio Avitabile for the helpful discussions; and we greatly appreciate the comments from two anonymous reviewers who helped improve the manuscript. This research is part of the Global Comparative Study (GCS) on REDD+ of the Center for International Forestry Research (CIFOR) (www.cifor.org/gcs) with funding support from the International Climate Initiative (IKI) of the German Federal Ministry for the Environment and Nuclear Safety (BMU), the Norwegian Agency for Development Cooperation, Australian Department of Foreign Affairs and Trade, European Commission (EC), UK Department for International Development, and the CGIAR Research Program on Forests, Trees and Agroforestry (CRP-FTA) with financial support from the CGIAR Fund Donors.

## Appendix ADecision trees for reclassification strategies


Map classes in original products
Map class in combined product
I Sensitive - early
>= 1 product change before
True
False
>= 1 product change after
2 products: no change
change before (1)
change after (2)
no change (3)
Adopt value of dataset that detects a disturbance the soonest, regardless of the other dataset's detection
II Sensitive - late
>= 1 product change after
>= 1 product change before
2 products: no change
change after (2)
change before (1)
no change (3)
Adopt value of dataset that detects a disturbance the latest, regardless of the other dataset's detection
III Conservative - early
>= 1 product no change
>= 1 product change before
2 products: change after
no change (3)
change before (1)
change after (2)
Any no change? Then adopt as no change.
Both change? Take earliest
IV Conservative - late
>= 1 product no change
>= 1 product change after
2 products: change before
no change (3)
change after (2)
change before (1)
Any no change? Then adopt as no change.
Both change? Take latest
V Timeless
Change before OR change after
>= 1 product change after
change (1)
no change (3)
For each individual product (GFC & BFAST):
Change in before or after? Then adopt as change.
No change? Then adopt as no change
Legend
1 Pixel value: Change in before period
2 Pixel value: Change in after period
3 Pixel value: No change
1 2 Pixel with conflicting values for the two products (one "change in before period", and other "change in after period")

After applying an overlay of the two map products, every pixel from the sample was reclassified according to the different reclassification strategies. The squares represent the possible pixel-level combinations of the two datasets. The rightmost column shows the decision in each of the reclassified map products.

## Appendix B Accuracies with 95%CI for all original and reclassified products

| | Product | Strata | UA | PA | OA | | Product | Strata | UA | PA | OA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Peru | GFC | Change before | 0.91 ± 0.08 | 0.40 ± 0.26 | 0.95 ± 0.04 | Indonesia-A | GFC | Change before | 0.85 ± 0.12 | 0.26 ± 0.11 | 0.85 ± 0.06 |
| | | Change after | 0.59 ± 0.11 | 0.73 ± 0.13 | | | | Change after | 0.79 ± 0.07 | 0.79 ± 0.19 | |
| | | No change | 0.96 ± 0.04 | 0.99 ± 0.00 | | | | No change | 0.87 ± 0.08 | 0.99 ± 0.01 | |
| | BFAST | Change before | 0.86 ± 0.08 | 0.49 ± 0.32 | 0.95 ± 0.04 | | BFAST | Change before | 0.80 ± 0.10 | 0.43 ± 0.17 | 0.83 ± 0.06 |
| | | Change after | 0.70 ± 0.12 | 0.54 ± 0.13 | | | | Change after | 0.76 ± 0.09 | 0.58 ± 0.15 | |
| | | No change | 0.96 ± 0.04 | 0.99 ± 0.00 | | | | No change | 0.84 ± 0.08 | 0.96 ± 0.02 | |
| | I | Change before | 0.84 ± 0.07 | 0.64 ± 0.40 | 0.96 ± 0.04 | | I | Change before | 0.78 ± 0.09 | 0.56 ± 0.21 | 0.88 ± 0.06 |
| | | Change after | 0.67 ± 0.11 | 0.94 ± 0.06 | | | | Change after | 0.76 ± 0.08 | 0.84 ± 0.20 | |
| | | No change | 0.98 ± 0.04 | 0.99 ± 0.00 | | | | No change | 0.91 ± 0.08 | 0.95 ± 0.02 | |
| | II | Change before | 0.84 ± 0.08 | 0.57 ± 0.36 | 0.96 ± 0.04 | | II | Change before | 0.77 ± 0.11 | 0.45 ± 0.17 | 0.86 ± 0.06 |
| | | Change after | 0.58 ± 0.10 | 0.97 ± 0.06 | | | | Change after | 0.70 ± 0.07 | 0.86 ± 0.20 | |
| | | No change | 0.98 ± 0.04 | 0.99 ± 0.00 | | | | No change | 0.91 ± 0.08 | 0.95 ± 0.02 | |
| | III | Change before | 0.97 ± 0.02 | 0.33 ± 0.21 | 0.94 ± 0.04 | | III | Change before | 0.91 ± 0.06 | 0.23 ± 0.09 | 0.82 ± 0.06 |
| | | Change after | 0.89 ± 0.09 | 0.30 ± 0.05 | | | | Change after | 0.96 ± 0.06 | 0.52 ± 0.13 | |
| | | No change | 0.94 ± 0.04 | 1.00 ± 0.00 | | | | No change | 0.80 ± 0.07 | 1.00 ± 0.00 | |
| | IV | Change before | 1.00 ± 0.00 | 0.26 ± 0.16 | 0.94 ± 0.04 | | IV | Change before | 1.00 ± 0.00 | 0.13 ± 0.05 | 0.81 ± 0.06 |
| | | Change after | 0.54 ± 0.07 | 0.33 ± 0.06 | | | | Change after | 0.80 ± 0.05 | 0.54 ± 0.13 | |
| | | No change | 0.94 ± 0.04 | 1.00 ± 0.00 | | | | No change | 0.80 ± 0.07 | 1.00 ± 0.00 | |
| | V-GFC | Change | 0.86 ± 0.06 | 0.54 ± 0.27 | 0.95 ± 0.04 | | V-GFC | Change | 0.96 ± 0.04 | 0.62 ± 0.14 | 0.88 ± 0.06 |
| | | No change | 0.96 ± 0.04 | 0.99 ± 0.00 | | | | No change | 0.87 ± 0.08 | 0.99 ± 0.01 | |
| | V-BFAST | Change | 0.89 ± 0.06 | 0.55 ± 0.27 | 0.96 ± 0.04 | | V-BFAST | Change | 0.84 ± 0.06 | 0.55 ± 0.13 | 0.84 ± 0.06 |
| | | No change | 0.96 ± 0.04 | 0.99 ± 0.00 | | | | No change | 0.84 ± 0.08 | 0.96 ± 0.02 | |
| Tanzania | GFC | Change before | 0.41 ± 0.15 | 0.05 ± 0.05 | 0.83 ± 0.08 | Indonesia-B | GFC | Change before | 0.97 ± 0.05 | 0.60 ± 0.08 | 0.94 ± 0.01 |
| | | Change after | 0.26 ± 0.15 | 0.10 ± 0.08 | | | | Change after | 0.74 ± 0.08 | 0.84 ± 0.09 | |
| | | No change | 0.86 ± 0.08 | 0.97 ± 0.01 | | | | No change | 0.96 ± 0.01 | 0.99 ± 0.01 | |
| | BFAST | Change before | 0.90 ± 0.10 | 0.14 ± 0.12 | 0.89 ± 0.08 | | BFAST | Change before | 0.79 ± 0.08 | 0.81 ± 0.08 | 0.93 ± 0.01 |
| | | Change after | 0.71 ± 0.15 | 0.50 ± 0.32 | | | | Change after | 0.92 ± 0.08 | 0.50 ± 0.09 | |
| | | No change | 0.90 ± 0.08 | 0.99 ± 0.01 | | | | No change | 0.94 ± 0.01 | 0.97 ± 0.01 | |
| | I | Change before | 0.67 ± 0.11 | 0.18 ± 0.15 | 0.87 ± 0.08 | | I | Change before | 0.81 ± 0.08 | 0.98 ± 0.03 | 0.97 ± 0.01 |
| | | Change after | 0.54 ± 0.14 | 0.54 ± 0.34 | | | | Change after | 0.86 ± 0.08 | 0.98 ± 0.02 | |
| | | No change | 0.91 ± 0.08 | 0.96 ± 0.01 | | | | No change | 1.00 ± 0.00 | 0.96 ± 0.01 | |
| | II | Change before | 0.67 ± 0.13 | 0.15 ± 0.13 | 0.87 ± 0.08 | | II | Change before | 0.80 ± 0.08 | 0.87 ± 0.03 | 0.96 ± 0.01 |
| | | Change after | 0.53 ± 0.14 | 0.54 ± 0.34 | | | | Change after | 0.76 ± 0.07 | 1.00 ± 0.00 | |
| | | No change | 0.91 ± 0.08 | 0.96 ± 0.01 | | | | No change | 1.00 ± 0.00 | 0.96 ± 0.01 | |
| | III | Change before | 0.77 ± 0.10 | 0.04 ± 0.03 | 0.85 ± 0.08 | | III | Change before | 0.98 ± 0.02 | 0.54 ± 0.05 | 0.91 ± 0.01 |
| | | Change after | 0.96 ± 0.06 | 0.06 ± 0.04 | | | | Change after | 0.96 ± 0.06 | 0.34 ± 0.03 | |
| | | No change | 0.85 ± 0.08 | 1.00 ± 0.00 | | | | No change | 0.91 ± 0.01 | 1.00 ± 0.00 | |
| | IV | Change before | 1.00 ± 0.00 | 0.01 ± 0.01 | 0.85 ± 0.08 | | IV | Change before | 1.00 ± 0.00 | 0.43 ± 0.04 | 0.90 ± 0.01 |
| | | Change after | 0.71 ± 0.06 | 0.06 ± 0.04 | | | | Change after | 0.66 ± 0.05 | 0.36 ± 0.04 | |
| | | No change | 0.85 ± 0.08 | 1.00 ± 0.00 | | | | No change | 0.91 ± 0.01 | 1.00 ± 0.00 | |
| | V-GFC | Change | 0.42 ± 0.16 | 0.11 ± 0.07 | 0.84 ± 0.08 | | V-GFC | Change | 0.94 ± 0.04 | 0.77 ± 0.06 | 0.95 ± 0.01 |
| | | No change | 0.86 ± 0.08 | 0.97 ± 0.01 | | | | No change | 0.96 ± 0.01 | 0.99 ± 0.01 | |
| | V-BFAST | Change | 0.91 ± 0.08 | 0.43 ± 0.22 | 0.90 ± 0.08 | | V-BFAST | Change | 0.84 ± 0.06 | 0.70 ± 0.06 | 0.93 ± 0.01 |
| | | No change | 0.90 ± 0.08 | 0.99 ± 0.01 | | | | No change | 0.94 ± 0.01 | 0.97 ± 0.01 | |
| Vietnam | GFC | Change before | 0.99 ± 0.03 | 0.29 ± 0.08 | 0.80 ± 0.06 | | | | | | |
| | | Change after | 0.68 ± 0.06 | 0.43 ± 0.16 | | | | | | | |
| | | No change | 0.81 ± 0.08 | 1.00 ± 0.00 | | | | | | | |
| | BFAST | Change before | 0.99 ± 0.02 | 0.73 ± 0.17 | 0.89 ± 0.06 | | | | | | |
| | | Change after | 0.88 ± 0.07 | 0.59 ± 0.21 | | | | | | | |
| | | No change | 0.88 ± 0.08 | 1.00 ± 0.01 | | | | | | | |
| | I | Change before | 0.98 ± 0.02 | 0.80 ± 0.18 | 0.92 ± 0.06 | | | | | | |
| | | Change after | 0.89 ± 0.07 | 0.68 ± 0.24 | | | | | | | |
| | | No change | 0.91 ± 0.08 | 1.00 ± 0.01 | | | | | | | |
| | II | Change before | 0.99 ± 0.03 | 0.62 ± 0.14 | 0.89 ± 0.06 | | | | | | |
| | | Change after | 0.73 ± 0.06 | 0.68 ± 0.24 | | | | | | | |
| | | No change | 0.91 ± 0.08 | 1.00 ± 0.01 | | | | | | | |
| | III | Change before | 0.99 ± 0.02 | 0.40 ± 0.09 | 0.80 ± 0.06 | | | | | | |
| | | Change after | 0.93 ± 0.07 | 0.33 ± 0.12 | | | | | | | |
| | | No change | 0.78 ± 0.07 | 1.00 ± 0.00 | | | | | | | |
| | IV | Change before | 1.00 ± 0.00 | 0.22 ± 0.05 | 0.78 ± 0.06 | | | | | | |
| | | Change after | 0.64 ± 0.05 | 0.34 ± 0.12 | | | | | | | |
| | | No change | 0.78 ± 0.07 | 1.00 ± 0.00 | | | | | | | |
| | V-GFC | Change | 1.00 ± 0.00 | 0.47 ± 0.10 | 0.84 ± 0.06 | | | | | | |
| | | No change | 0.81 ± 0.08 | 1.00 ± 0.00 | | | | | | | |
| | V-BFAST | Change | 0.99 ± 0.02 | 0.69 ± 0.15 | 0.90 ± 0.06 | | | | | | |
| | | No change | 0.88 ± 0.08 | 1.00 ± 0.01 | | | | | | | |

## Appendix C Paired T-tests

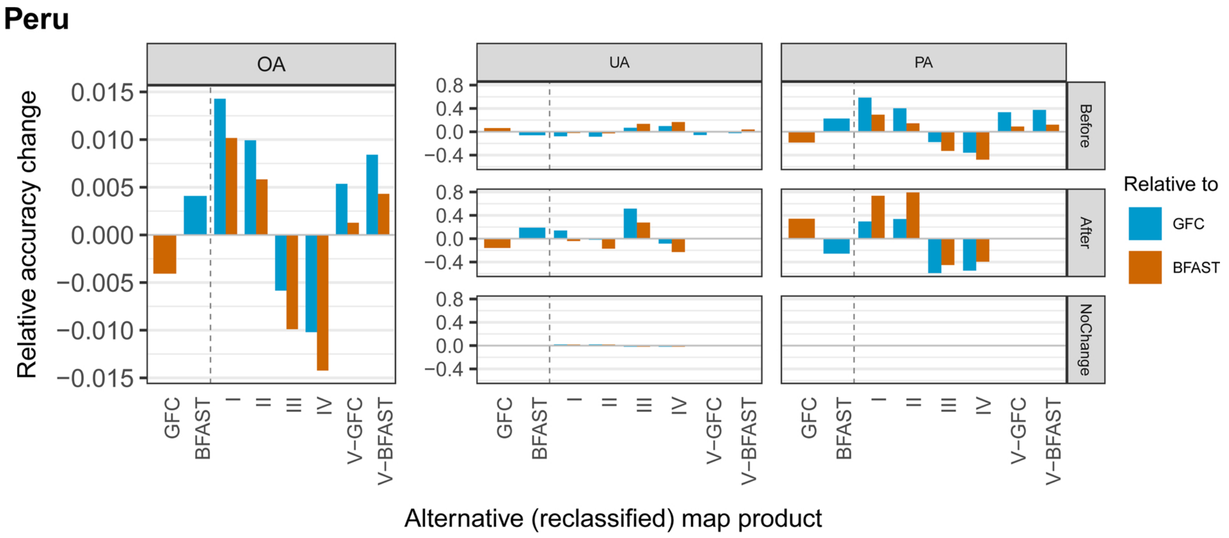
| | (reclassified) product | mean | median | sd | Mean Δ | p-value | Mean Δ | p-value |
|---|---|---|---|---|---|---|---|---|
| | GFC | 0.88 | 0.85 | 0.07 | NA | NA | −0.02 | 0.82 |
| | BFAST | 0.90 | 0.89 | 0.05 | 0.02 | 0.18 | NA | NA |
| Overall Accuracy | I | 0.92 | 0.92 | 0.05 | 0.04** | 0.04 | 0.02* | 0.08 |
| | II | 0.91 | 0.89 | 0.05 | 0.03** | 0.05 | 0.01 | 0.19 |
| | III | 0.87 | 0.85 | 0.06 | −0.01 | 0.81 | −0.03 | 0.95 |
| | IV | 0.86 | 0.85 | 0.07 | −0.02 | 0.92 | −0.04 | 0.96 |
| | V – GFC | 0.91 | 0.90 | 0.04 | 0.03 | 0.13 | 0.01*** | 0.01 |
| | V – BFAST | 0.89 | 0.88 | 0.06 | 0.02** | 0.02 | −0.01 | 0.59 |

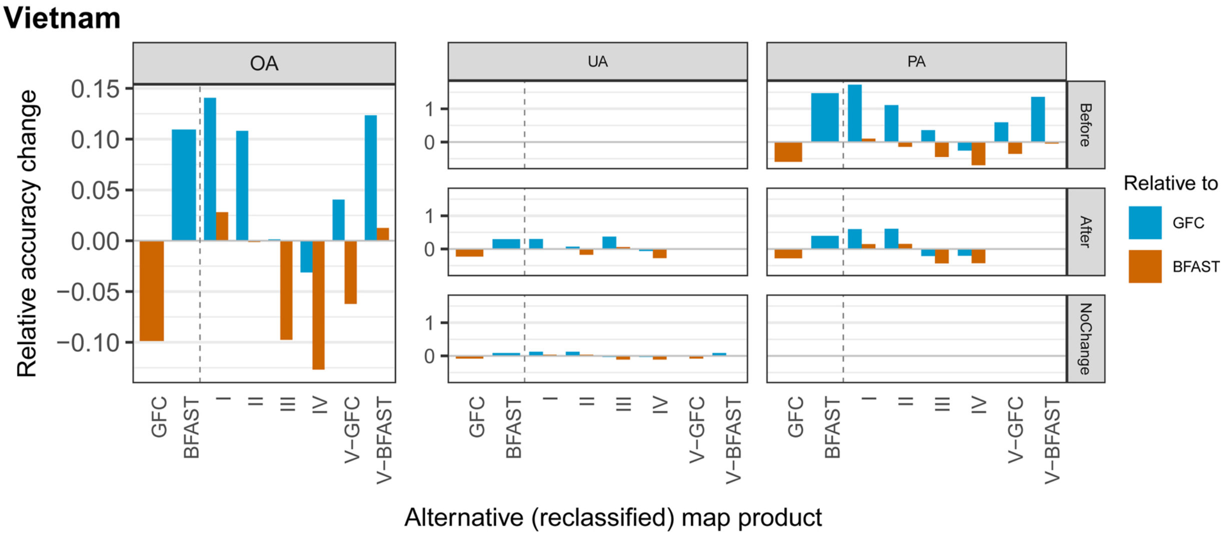
| USER'S ACCURACY | | | | | BEFORE | | | | | AFTER | | | | | NO CHANGE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Relative to GFC | Relative to BFAST | | | | Relative to GFC | Relative to BFAST | | | | Relative to GFC | Relative to BFAST |
| (reclassified) product | mean | median | sd | | Mean Δ | Mean Δ | mean | median | sd | Mean Δ | Mean Δ | mean | median | sd | Mean Δ | Mean Δ |
| GFC | 0.82 | 0.91 | 0.24 | | NA | −0.04 | 0.61 | 0.68 | 0.21 | NA | −0.18 | 0.89 | 0.87 | 0.07 | NA | −0.02 |
| BFAST | 0.87 | 0.86 | 0.08 | | 0.04 | NA | 0.79 | 0.76 | 0.10 | 0.18** | NA | 0.91 | 0.90 | 0.05 | 0.02 | NA |
| I | 0.82 | 0.81 | 0.11 | | −0.01 | −0.05 | 0.74 | 0.76 | 0.14 | 0.13** | −0.05 | 0.94 | 0.91 | 0.04 | 0.05*** | 0.04** |
| II | 0.81 | 0.80 | 0.12 | | −0.01 | −0.06 | 0.66 | 0.70 | 0.10 | 0.05 | −0.14 | 0.94 | 0.91 | 0.04 | 0.05*** | 0.04** |
| III | 0.92 | 0.97 | 0.09 | | 0.10 | 0.06 | 0.94 | 0.96 | 0.03 | 0.33** | 0.14** | 0.86 | 0.85 | 0.07 | −0.03 | −0.05 |
| IV | 1.00 | 1.00 | 0.00 | | 0.18* | 0.13** | 0.67 | 0.66 | 0.10 | 0.06 | −0.12 | 0.86 | 0.85 | 0.07 | −0.03 | −0.05 |
| V – GFC | 0.83 | 0.94 | 0.24 | | 0.01 | −0.03 | NA | NA | NA | NA | NA | 0.89 | 0.87 | 0.07 | NA | −0.02 |
| V – BFAST | 0.89 | 0.89 | 0.06 | | 0.07 | 0.03** | NA | NA | NA | NA | NA | 0.91 | 0.90 | 0.05 | 0.02 | NA |

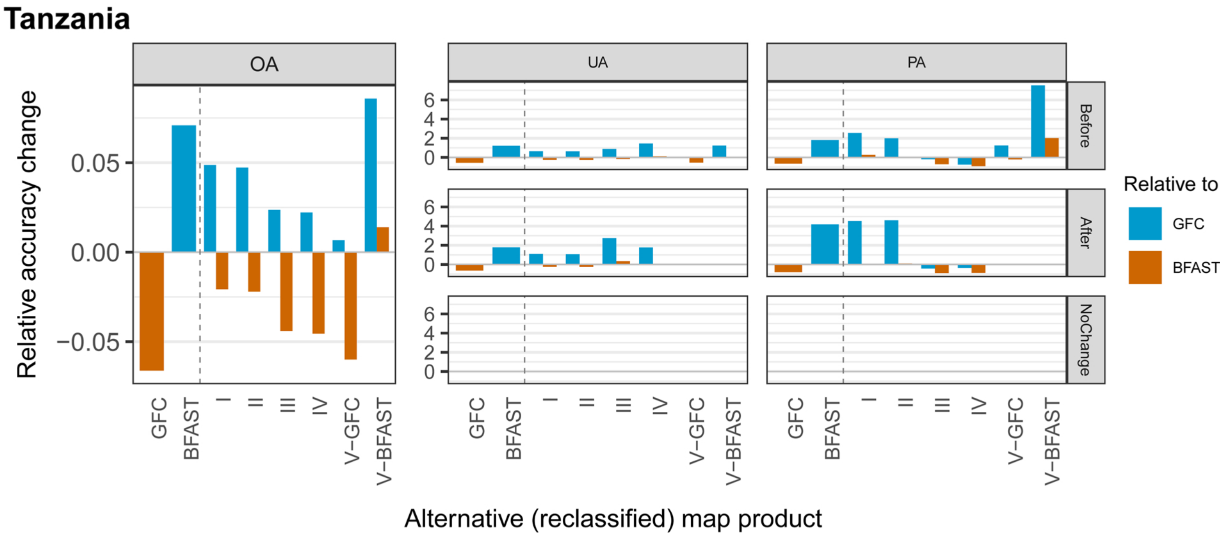
| PRODUCER'S ACCURACY | | | | | BEFORE | | | | | AFTER | | | | | NO CHANGE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Relative to GFC | Relative to BFAST | | | | Relative to GFC | Relative to BFAST | | | | Relative to GFC | Relative to BFAST |
| (reclassified) product | mean | median | sd | | Mean Δ | Mean Δ | mean | median | sd | Mean Δ | Mean Δ | mean | median | sd | Mean Δ | Mean Δ |
| GFC | 0.32 | 0.29 | 0.20 | | NA | −0.20 | 0.58 | 0.73 | 0.31 | NA | 0.03 | 0.99 | 0.99 | 0.01 | NA | 0.01 |
| BFAST | 0.52 | 0.49 | 0.26 | | 0.20** | 0.00 | 0.54 | 0.54 | 0.05 | −0.03 | NA | 0.98 | 0.99 | 0.02 | −0.01 | NA |
| I | 0.63 | 0.64 | 0.30 | | 0.31*** | 0.11*** | 0.79 | 0.84 | 0.18 | 0.22** | 0.25** | 0.97 | 0.96 | 0.02 | −0.02 | −0.01 |
| II | 0.53 | 0.57 | 0.26 | | 0.21*** | 0.01 | 0.81 | 0.86 | 0.19 | 0.24*** | 0.27** | 0.97 | 0.96 | 0.02 | −0.02 | −0.01 |
| III | 0.31 | 0.33 | 0.19 | | −0.01 | −0.21 | 0.31 | 0.33 | 0.17 | −0.27 | −0.24 | 1.00 | 1.00 | 0.00 | 0.01** | 0.02** |
| IV | 0.21 | 0.22 | 0.15 | | −0.11 | −0.31 | 0.33 | 0.34 | 0.17 | −0.25 | −0.22 | 1.00 | 1.00 | 0.00 | 0.01** | 0.02** |
| V – GFC | 0.50 | 0.54 | 0.24 | | 0.18** | −0.02 | NA | NA | NA | NA | NA | 0.99 | 0.99 | 0.01 | NA | 0.01 |
| V – BFAST | 0.58 | 0.55 | 0.11 | | 0.26*** | 0.07 | NA | NA | NA | NA | NA | 0.98 | 0.99 | 0.02 | −0.01 | NA |

$H_0$: means do not differ.
$H_a$: means of reclassification product > input product.
N = 5 (sites).
*significance level 0.90.
**significance level 0.95.
*** significance level 0.99.

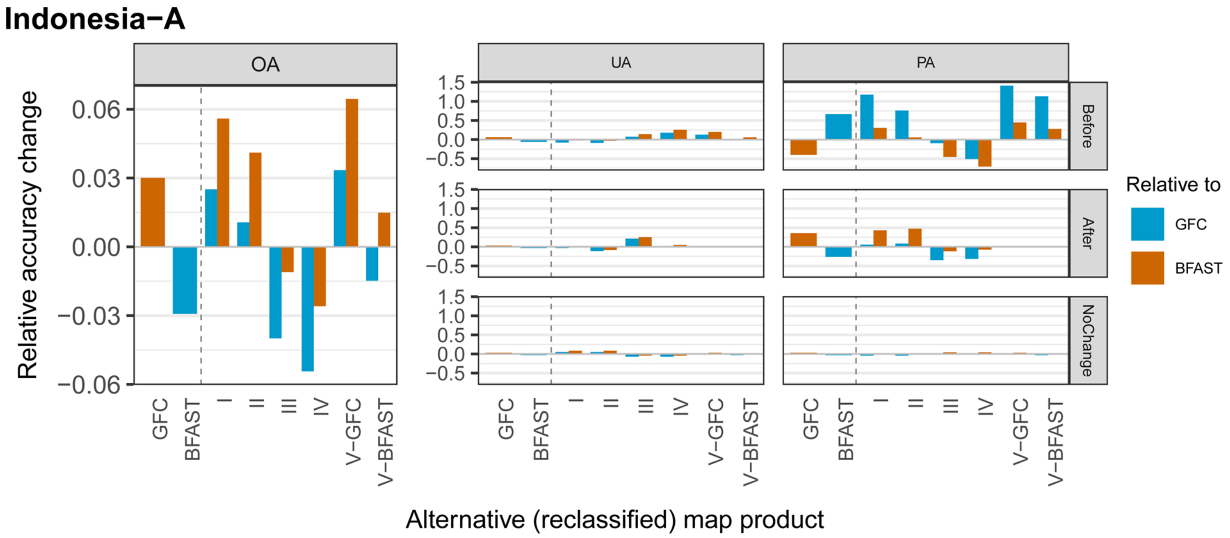
## Appendix D. Relative Accuracy Changes Per Site

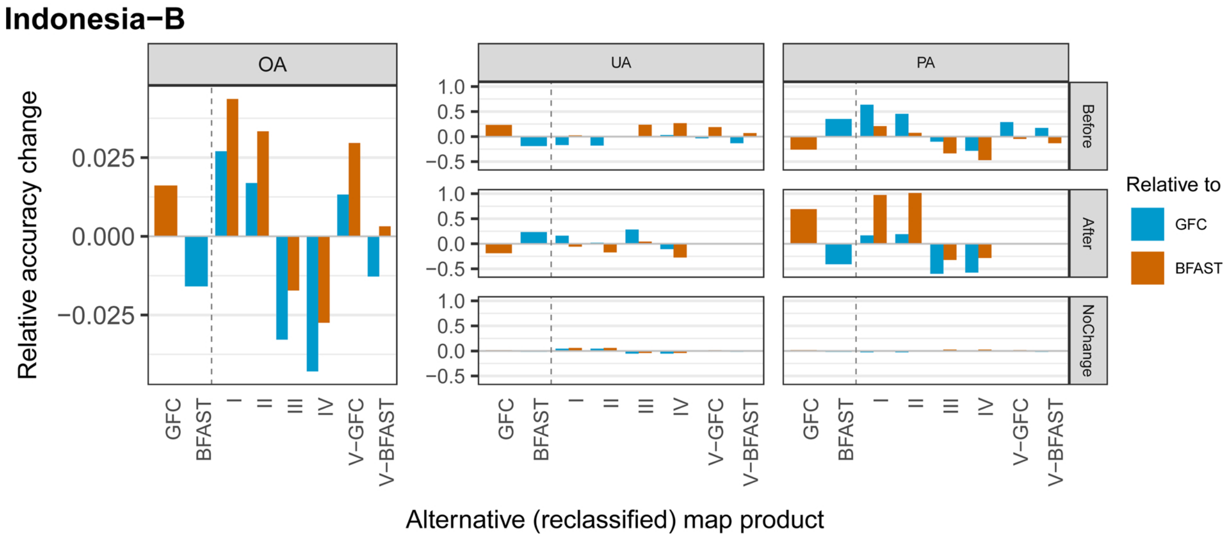






**Appendix D. (continued)**





## References

Bos, A.B., Duchelle, A.E., Angelsen, A., Avitabile, V., Sy, V., De, Herold, M., Joseph, S., Sassi, Cde, Sills, E.O., Sun derlin, W.D., Wun der, S., 2017. Comparing methods for assessing the effectiveness of subnational REDD+ initiatives. Environ. Res. Lett. 12, 074007. https://doi.org/10.1088/1748-9326/aa7032.

CIFOR, 2017. Global Comparative Study on REDD+ Subnational REDD+ Initiatives. [WWW Document]. URL https://www.cifor.org/gcs/modules/redd-subnational-initiatives/ (accessed 2.1.19). .

Cochran, W.G., 1977. Sampling Techniques. Wiley, New York.

Cohen, W.B., Yang, Z., Kennedy, R., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — Tools for calibration and validation. Remote Sens. Environ. 114, 2911–2924. https://doi.org/10.1016/j.rse.2010.07.010.

DeVries, B., Decuyper, M., Verbesselt, J., Zeileis, A., Herold, M., Joseph, S., 2015. Tracking disturbance-regrowth dynamics in tropical forests using structural change detection and Landsat time series. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2015.08.020.

Duchelle, A.E., Herold, M., de Sassi, C., 2015. Monitoring REDD+ Impacts: Cross Scale Coordination and Interdisciplinary Integration, in: Sustainability Indicators in Practice. De Gruyter Open, Berlin, Germany, pp. 55–79. https://doi.org/10.1515/9783110450507-009.

Foody, G.M., 2009. Sample size determination for image classification accuracy assessment and comparison. Int. J. Remote Sens. 30, 5273–5291. https://doi.org/10.1080/01431160903130937.

GFOI, 2016. Integration of Remote-sensing and Ground-based Observations for Estimation of Emissions and Removals of Greenhouse Gases in Forests (No. Edition 2.0). Rome.

GOFC-GOLD, 2016. A Sourcebook of Methods and Procedures for Monitoring and Reporting Anthropogenic Greenhouse Gas Emissions and Removals Associated With Deforestation, Gains and Losses of Carbon Stocks in Forests Remaining Forests, and Forestation (No. COP22 Version 1), GOFC-GOLD Report, GOFC-GOLD Report. Wageningen University, Wageningen, the Netherlands.

Grassi, G., Monni, S., Federici, S., Achard, F., Mollicone, D., 2008. Applying the conservativeness principle to REDD to deal with the uncertainties of the estimates. Environ. Res. Lett. 3, 035005. https://doi.org/10.1088/1748-9326/3/3/035005.

Grassi, G., House, J., Dentener, F., Federici, S., den Elzen, M., Penman, J., 2017. The key role of forests in meeting climate targets requires science for credible mitigation. Nat. Clim. Change 7. doi:10.0.4.14/nclimate3227.

Gross, D., Achard, F., Dubois, G., Brink, A., Prins, H.H.T., 2017. Uncertainties in tree cover maps of Sub-Saharan Africa and their implications for measuring progress towards CBD Aichi Targets. Remote Sens. Ecol. Conserv. 1–19. https://doi.org/10.1002/rse2.52.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 80 (342), 850–853. https://doi.org/10.1126/science.1244693.

IPCC, 2006. Generic methodologies applicable to multiple land-use categories. In: Eggleston, S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K. (Eds.), IPCC Guidelines for National Greenhouse Gas Inventories - Vol 4 AFOLU. Institute for Global Environmental Strategies (IGES), Hayama, Japan, Hayama, Japan, pp. 2.1–2.59.

Melo, J.B., Ziv, G., Baker, T.R., Carreiras, J.M.B., Pearson, T.R.H., Vasconcelos, M.J.,

2018. Striking divergences in Earth Observation products may limit their use for REDD+. Environ. Res. Lett. 13, 104020. https://doi.org/10.1088/1748-9326/aae3f8.

Milodowski, D.T., Mitchard, E.T.A., Williams, M., 2017. Forest loss maps from regional satellite monitoring systematically underestimate deforestation in two rapidly changing parts of the Amazon. Environ. Res. Lett. 12, 094003. https://doi.org/10.1088/1748-9326/aa7e1e.

Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sens. Environ. 129, 122–131. https://doi.org/10.1016/j.rse.2012.10.031.

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

Open Foris, 2019. Open Foris - Free Open-source Solutions for Environmental Monitoring. [WWW Document]. URL http://www.openforis.org (accessed 2.1.19). .

Romijn, E., Lantican, C.B., Herold, M., Lindquist, E., Ochieng, R., Wijaya, A., Murdiyarso, D., Verchot, L., 2015. Assessing change in national forest monitoring capacities of 99 tropical countries. For. Ecol. Manage. 352, 109–123. https://doi.org/10.1016/j.foreco.2015.06.003.

Sills, E.O., Atmadja, S., Sassi, C., de, Duchelle, A.E., Kweka, D., Resosudarmo, I.A.P., Sun derlin, W.D., 2014. REDD+ On the Ground: a Case Book of Subnational Initiatives across the Globe. Center for International Forestry Research (CIFOR), Bogor, Indonesia. https://doi.org/10.17528/cifor/005202.

Simonet, G., Karsenty, A., Perthuis, C., de, Newton, P., Schaap, B., 2015. REDD+ Projects in 2014: an Overview Based on a New Database and Typology (No. 32), Les Cahiers De La Chaire Economie Du Climat - Information and Debates Series, Les Cahiers De La Chaire Economie Du Climat Information - Information and Debates Series. Chaire Economie du Climat, Paris, France.

Stehman, S.V., 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. Int. J. Remote Sens. 35, 4923–4939. https://doi.org/10.1080/01431161.2014.930207.

Sunderlin, W.D., Sills, E.O., Duchelle, A.E., Ekaputri, A.D., Kweka, D., Toniolo, M.A., Ball, S., Doggart, N., Pratama, C.D., Padilla, J.T., Enright, A., Otsyina, R.M., 2015. REDD+ at a critical juncture: assessing the limits of polycentric governance for achieving climate change mitigation. Int. For. Rev. 17, 400–413. https://doi.org/10.1505/146554815817476468.

Turubanova, S., Potapov, P.V., Tyukavina, A., Hansen, M.C., 2018. Ongoing primary forest loss in Brazil, Democratic Republic of the Congo, and Indonesia. Environ. Res. Lett. 13, 074028. https://doi.org/10.1088/1748-9326/aacd1c.

UNFCCC, 2006. Good Practice Guidance and Adjustments Under Article 5, Paragraph 2, of the Kyoto Protocol.

UNFCCC, 2019. REDD+ Web Platform - Submissions. (Accessed 8.23.18)([WWW Document]. URL). https://redd.unfccc.int/submissions.html.

Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sens. Environ. 114, 106–115. https://doi.org/10.1016/j.rse.2009.08.014.

Verbesselt, J., Zeileis, A., Herold, M., 2012. Near real-time disturbance detection using satellite image time series. Remote Sens. Environ. 123, 98–108. https://doi.org/10.1016/j.rse.2012.02.022.

Verchot, L.V.V., Anitha, K., Romijn, E., Herold, M., Hergoualc'h, K., 2012. Emissions factors: converting land use change to CO2 estimates. In: Angelsen, A. (Ed.), Analysing REDD+: Challenges and Choices. Center for International Forestry Research (CIFOR), Bogor, Indonesia, pp. 261–278.

Wong, G., Angelsen, A., Brockhaus, M., Carmenta, R., Duchelle, A.E., Leonard, S., Luttrell, C., Martius, C., Wunder, S., 2016. Results-based payments for REDD+: lessons on finance, performance, and non-carbon benefits. CIFOR Infobr. https://doi.org/10.17528/cifor/006108.