##

Anonymous Referee #1

Received and published: 19 February 2021

This manuscript presents a new burned area product based on Sentinel-3 data by adapting the burned area mapping algorithm used for the FireCCI51 product that is based on BODIS data, to the new Sentinel-3 data.

To follow the manuscript, especially due to the way it is written, a pre-knowledge about these algorithms from the previous papers would be desirable. The authors many times in the manuscript they describe the algorithms given that the reader is aware of these algorithms.

Although the new algorithm is adapted to the Sentiel-3 data, it makes extensive use of MODIS data by using thermal anomalies, therefore for this point of view the new algorithm may not be considered that it is one exclusively referring to the Sentinel-3 data.

In the Introduction the authors put a lot of emphasis on FireCCI project, from one point of view is justified, but I think it is over-discussed.

I will not make any comment to the algorithm since this is already published (the main core) and also because it is not that much explained in the manuscript, but I will focus my comments on the accuracy of the new product.

The evaluation of the new product is implemented at multiple levels, but I think the most important should be the paragraph mentioned as spatial assessment. The authors also evaluate the product using a consistency assessment with other products, but in this case it is expected to be high since there are common data and common methods between the products. Coming back to the spatial assessment as the author name this, I would like to see a more in depth evaluation and discussion. First of all, the omission is almost 50% (which means that half of the fires are missed) and there is an additional 20% of commission errors which means that the overall error is quite high. For this, I would expect a more in depth evaluation and discussion of the errors. Also, I propose the authors to make an assessment of the errors against the uncertainty assessment. For example, to estimate the accuracy of the new product for different levels of uncertainty assessment. Additionally to this, I would like to see an additional exploration of the spatial and maybe temporal patterns of the errors. In other words are there patterns in the spatial or temporal dimension for the errors?

##

Anonymous Referee #2

Received and published: 8 March 2021

General Comments:

This manuscript presents the methods and results for creating an operational Sentinel-3- (OLCI) and MODIS-derived burned area product created under the Climate Change Initiative and Copernicus Climate Change Service. The algorithm is an adaptation of the MODIS-based FireCCI51 product. The manuscript's organization is logical, but it requires minor proofreading throughout for both grammar and typography (e.g., consistency with units like km2) that goes

beyond the scope of peer review. In general, many methods in the paper refer to external works that reduce the transparency of this work. Particularly, the discussion of the validation methods is far too vague to be understood by the reader, especially considering the study does not implement the short sampling units of the work that it cites. Additionally, the lack of discussion of the dataset's limitations is problematic for a data descriptor manuscript, especially for burned area products that typically demonstrate high omission error rates.

There are several major criticisms of this work as a whole. First and foremost, the C3SBA10 product is not an improvement over its predecessor when considering the accuracy metrics presented by the authors (Dice coefficient is lower in all cases, commission errors are slightly better or the same, omission errors are higher in all cases, and relative bias is higher in all cases). Simultaneously, its dependence on the MODIS sensor means that the product cannot be generated any further into the future than the FireCCI51 product that it replaces. Thus, when the MODIS sensors are decommissioned ca. 2023, C3SBA10 will no longer be functional in the proposed form (and presumably be replaced by another version of the product using a different active fire data source, that will also be submitted for peer review in another manuscript). So, the authors' assertion on line 98 that this work will "ensure that multi-decadal analyses can benefit from both datasets uninterruptedly" is untrue and unrealistic as C3SBA10 can never be multi-decadal. C3SBA10, therefore, does not improve the quality of the available burned area data, nor does it extend the length of the burned area archive, making its value very limited.

The presentation of the results and validation raise many flags as well. The authors present the relationship between FireCCI51 and C3SBA10 as a function of the linear regression's slope. The reader is left to assume that the simple linear regression (ordinary least squares) was used – this is not the correct method because OLS regression assumes the independent variable is free of errors, and the choice of axis will change the result (i.e., making C3SBA the independent variable and FireCCI51 the dependent variable will yield a completely different result). A Deming regression, like total least squares, is more appropriate as it does not assume an error-free independent variable, nor does it assume variable dependence.

The comparison of RMSE and regression slopes are in a gray area given that the products are created using essentially the same algorithm and data that is as similar as possible. Both results could be completely wrong, for the same reason, and show high agreement. Comparison to an independent data source would be more informative

There is no acknowledgment of any of the proposed methods' limitations and resulting product throughout the work. For example, the dependence on MODIS is a substantial drawback of this method that cannot be easily remedied using alternative data sources (discussed at the end of this review). The C3SBA10 product itself shows lower accuracy than the FireCCI51 product that it replaces, but consistency comparisons are emphasized instead. In general, the fire community should aim to make more accurate products, not less accurate products, and the continued generation of products that perform worse than their predecessors serves only to erode users' confidence in the products. The authors have taken an incremental improvement and versioning approach with the previous FireCCI50 and FireCCI51 products. However, a key distinction is that the FireCCI50 method was novel, and FireCCI51 made a material change to the algorithm and showed an improvement in accuracy – C3SBA10, on the other hand, does not use a novel method and does not improve on previous results. If the C3SBA10 product did improve (significantly) upon the accuracy of the FireCCI51 product, or implemented an active fire product with potential for long-term use, the manuscript would be more compelling.

Specific Comments:

The manuscript's introduction includes an overview of the FireCCI, C3S, and ECV's that goes far beyond what is relevant to the reader – a simpler summary and reference would suffice. Similarly, the discussion of technical specifications of OLCI in section 2.2.1 is rather extensive, but the sensor specifications are not discussed in the context of the results. Even though the results section attributes any discrepancies to sensor differences, little effort is made to identify what properties are responsible for specific discrepancies.

In section 2.4, the authors mention that the product is produced in 10x10 degree tiles like the predecessor product. A recent paper by Liu and Crowley (2021) identifies what were described as "severe tiling artifacts" in the FireCCI51 product. Is this also the case in the C3SBA10 product?

Section 2.6.2 notes that active fire data was used to assess the temporal accuracy of the product. The reader would assume this means the MCD14ML product – but the manuscript does not establish how this can be considered independent of the burned area product generation itself, which is very important given that MCD14ML is a direct input into the burned area product. Why was an independent sensor like VIIRS or SLSTR not used to avoid this issue altogether? Noteworthy – the method that the authors replicate (Boschetti et al., 2010) for the temporal accuracy assessment was designed for the MCD45A1 burned area product that did not use active fire detections as an input.

In section 2.7, the authors note that "C3SBA10 operational product cannot be understood as a unique, independent dataset, but as a continuation of its predecessor FireCCI51" – why not compare to an independent product like MCD64A1?

Section 3 – A 3 year-trend is not a significant trend, as noted by the authors; why not just call this the summary of the annual burned area? With no consideration for fire seasonality, how can the authors be certain that the "trend" is meaningful rather than a slightly shifted fire season in Africa from December through January?

In line 343, it is noted that almost all burned area in tropical forests occurred between 20 N and 20 S – this is, by definition, the only latitude band where tropical forests exist, so this observation is meaningless.

Section 3.2 (and subsequent relevant sections) – The manuscript barely discusses the causes of errors in the product and spends significantly more time on the consistency assessment (comparing the algorithm to itself) than on accuracy assessment (comparing to independent data). The emphasis placed on consistency assessment, combined with the "long units" accuracy assessment, is part of the broader pattern of de-emphasizing any negative aspect of the work and re-framing it in a positive light. There should be a critical and realistic self-evaluation of the product so that users can understand its limitations.

In line 382, do the authors mean "significantly higher for 2019" in the context of statistical significance?

In lines 383-384, the authors state that "more images due to the presence of two S3 satellites can explain why 2019 was the most similar year," but do not offer any proof for the statement. The algorithm could be run with only S3A as a comparison to test the assertion. Additionally, while the result may be more similar to the FireCCI51 product, it was also the worst of the three years for commission and omission error – why does doubling the number of observations lead to a worse result?

Section 3.2.2. – 10 days is an extremely long detection window following active fire detection. The Boschetti et al. (2010) paper, which the authors referenced in the manuscript, found that the MCD45A1 product was vastly superior in preserving fire timing more than ten years ago (75% of burn dates were within 4 days of the active fire detection, vs. 64.6% within 5 days for the present manuscript). Given that this is a dataset description manuscript, users would benefit from an explanation of why the burn detection dates are so far behind the actual day of burning.

Section 3.3 – It seems as though the underestimation of C3SBA10, relative to FireCCI51, is being "obscured" in how statistics are presented here. For example, the 0.65, 0.56, and 0.28 Mkm^2 differences noted in the manuscript correspond to differences of approximately 17%, 15.5%, and 8%, respectively (percentages not included in the manuscript). These are large deficits, regardless of RMSE, regression coefficients, and slope, making it hard to believe that the results truly are consistent. The use of RMSE reported in square kilometers does not make sense because the analysis grid is in degrees. The maximum error at the poles approaches zero, while the maximum error at the Equator is approximately 12,000 sq km.

Line 452 – I don't think the word "traduced" is being used appropriately.

Lines 495-496 – The discussion of long vs. short units highlights why the use of long units is inappropriate. The manuscript states: "The short units' approach is more affected by the temporal reporting accuracy of the global BA products than when using long units." This shows that the method is fitting the validation data and methods to match the result. At a fundamental level, there is no certainty that the fire observed in the long sampling unit is the same fire that was observed in the burned area dataset, given that the fire observations can be up to a year apart (as described in 2.6.1). This is especially the case in very fire-prone environments like African Savannahs that may burn more than once per year, and minimizing the time between validation scene observations is done to avoid misrepresenting these errors. By the authors' own writing, the validation method was purposefully constructed to accommodate the temporal errors in the dataset (i.e., recast incorrect classifications as correct classifications).

Lines 524-528 – The discussion about human impacts on fires does not seem relevant to this work.

Lines 529-537 – This paragraph's thesis states that sensor characteristics affect the products, but then the following sentences don't support that thesis concretely. The authors should provide evidence for causation here, the discussion of croplands and shrublands is anecdotal.

Discussion – If this work's goal is to continue the time series for multi-decadal analysis, it makes no sense to rely on the same sensor used in the predecessor work because the time series will effectively be ended once MODIS is decommissioned. The authors note that VIIRS could be a replacement, but this is only partially true because VIIRS does not have a morning overpass, so the effective temporal resolution is worse than MODIS. The authors reference the SLSTR algorithm (Xu et al., 2020) as having good capabilities for small fires, but the algorithm they referenced is currently nighttime-only. SLSTR is poorly suited for daytime detections because of the mid-infrared bands' saturation over surfaces hotter than 38C – is there any indication from the SLSTR algorithm developers that a viable/effective daytime algorithm will be available soon and if so, how would that affect the implementation of the present algorithm? Given both the limitations and longevity of VIIRS and SLSTR, it would make sense for this work to implement or test either of those sensors as they represent the only paths forward in a post-MODIS era. There are, therefore, significant challenges associated with this product in the near future that need to be addressed.

References:

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(6), 705-709.

Liu, T., & Crowley, M. A. (2021). Detection and impacts of tiling artifacts in MODIS burned area classification. IOP SciNotes, 2(1), 014003.

Xu, W., Wooster, M. J., He, J., & Zhang, T. (2020). First study of Sentinel-3 SLSTR active fire detection and FRP retrieval: Night-time algorithm enhancements and global intercomparison to MODIS and VIIRS AF products. Remote Sensing of Environment, 248, 111947.

##

Answer to the reviewers

As suggested by the EiC of this manuscript, we are replying the reviewers' comments in general terms, indicating which are the main novelties of the new version of the manuscript, which includes all relevant suggestions and comments raised by the reviewers. We appreciate their effort to provide an in-depth review of our paper.

First, we should clarify that the goal of the paper was not to present a new product, but rather an extension of a previous product using a new sensor, which will be used as part of an operational climate service. Therefore, the critical issue was to adapt the precursor algorithm (FireCCI51) to a new sensor and demonstrate that performance was similar as to derive a multidecadal time series of BA data (from 2001 to the present). The use of MODIS active fires will be certainly a limitation of this new operational product in the near future, but their replacement with either VIIRS or SLSTR data is not critical for the continuation of the product. We have in fact tested this issue in several tiles and demonstrate that similar BA results can derived from VIIRS.

The reviewers' comments indicate that the paper is not very novel as it does not provide a new algorithm, but this was not the intention of this paper, but rather to show the consistency of a long-term product, created by two different sensors. This is not a trivial task, as the reviewers can check in other ECVs (Hsu et al., 2019;Sayer et al., 2018;Sayer et al., 2017;Merchant et al., 2020). The C3S programme is operational service. It assumes that products are mature enough to be produced routinely, but this is not the actual case of BA data which still require a lot of research to reduce the omission and commission errors observed in existing global products (not just our datasets but also NASA's, with similar accuracy values to ours). Since BA information is required by climate modellers, a compromise is to continue working to improve observed problems, while simultaneously providing the community with the information necessary to use appropriately the existing products. For this reason, all BA products generated within the FireCCI project have included a global validation analysis, simultaneous to the product release. Unfortunately, this has not been the case of other existing BA products, leading to their overuse, sometimes with users not being fully aware of their limitations. We have also included a validation for the C3SBA10 product in this manuscript, providing the relevant accuracy values. A more detailed analysis of the validation effort, exceeding the focus on this manuscript, was submitted to RSE and it is now in the review process. In any case, we agree with the reviewers that the previous version of the manuscript did not properly reflect the product limitations and

therefore we have extended these comments in the discussion section. We now comment about the sources of errors, including those leading to temporal reporting accuracy and potential border effects.

The C3S service relies on transferring the knowledge generated from research projects such as the ESA CCI. Part of this transferring regards the adaption of previous algorithms to the new Copernicus Sentinel missions, again with the emphasis of obtain a consistent product for long-term analysis. In the case of the BA products, it was decided to make a conservative choice by transferring the experience from the latest version of FireCCI algorithms (FireCCI51), which was based on MODIS, to the OLCI sensor on board Sentinel-3 satellites, as both have similar spatial resolution and common spectral bands. Since the FireCCI51 product was conceived as a research product, not an operational one, it was processed only until 2019. Therefore, the new version of the product would guarantee the continuity of the CCI BA time series from 2019 to, at least, 2030. Since consistency was critical, we have emphasised in the manuscript the inter-comparison analysis, to show that spatial and temporal trends are similar enough to use the resulting time series reliably. We have redone the correlation analysis, as suggested by one reviewer, using TLS, obtaining similar results to the previous version. Regarding RMSE errors, we were already aware about the correction of grid size with latitude, and for this reason they are expressed in km2. We have also shown that the C3SBA10 and its precursor show more similar values when the two S-3 satellites were operating, which was expected. Accuracy measures were also more similar than in the two previous years.

Regarding the source of the active fires, we should indicate that they are an auxiliary dataset that can be replaced in the near future, by other sensors. In that sense, we consider two possibilities: the SLSTR sensor on board Sentinel-3 or the VIIRS sensor on board Suomi-NPP and NOAA-20. The former case provides a unique opportunity to generate a global BA product fully based on European sensors. However, as the reviewers correctly pointed out, for the moment, although it showed a good performance, the active fire algorithm of SLSTR is night time-only. This leads to a partial monitoring of the actual global fire activity. Prof. Wooster (personal communication) expects to release a day-time version of his algorithm by the beginning of 2022, well before the end of the MODIS acquisitions, so this would be a real possibility for replacement, obviously if the quality is high enough. In the case of the VIIRS sensor, the active fire product has shown a substantial increase in detections of small fires due to its improved 375 m spatial resolution (versus 1 km of MODIS). However, the lack of a morning overpass could be a limitation considering the diurnal cycle of tropical fires. Still, to show how this replacement may impact the BA detections in the future, we processed our BA algorithm for the 13 calibration tiles for the year 2019 replacing MODIS with VIIRS active fires (VNP14IMGML C1 product). The results were compared to 43 Landsat reference perimeters that were additionally interpreted for the 13 study sites. The results showed that the C3SBA10 performance is similar with both active fire datasets (Table 1). In fact, the results slightly improved. Therefore, this preliminary exercise demonstrates the viability of replacing one active fire product by other in the C3SBA10 algorithm with similar results. This analysis was not included in the paper because we consider it out of scope.

Table 1. Accuracy metrics for test tiles.

Metric	C3SBA10 with MODIS	C3SBA10 with VIIRS
Dice coefficient	68.1	71.8

Commission error	20.5	22.3
Omission error	40.5	33.3
Relative bias	-25.1	-14.1

Regarding the spatial validation, the results showed in this paper are part of another manuscript that is being peer-reviewed in RSE (Franquesa et al., 2021). In that paper, the authors show how the temporal reporting accuracy errors of the BA products directly affect commission and omission errors. However, these reporting accuracy errors must be understood as those pixels that are correctly labelled as burned in spatial terms but suffer from a temporal delay in the detection. Therefore, this type of temporal errors become a spatial error if the burned pixel is detected by the BA product outside the reference period. The shorter that period, the higher the probability of an actual burned pixel be reported outside the validation period (and therefore labelled as a detection error). This affects all BA products (as shown in Franquesa et al. (2021) paper), not just ours, but certainly it has a higher implication for those products that have a lower temporal reporting accuracy. Ours indeed has this limitation (as indicated in the manuscript), and therefore the comparison with short-time reference units affects even more than to other BA products.

We consider conceptually different both errors, which are commonly mixed in validation exercises, as it is further discussed in the Franquesa et al.'s paper (see figure 1 below extracted from this manuscript to see a clear example on the impact of reporting accuracy on spatial accuracy). The drawbacks of long reference units are also analysed in this paper. Regarding the potential problem burns of early and late season that may occur in frequently burned areas, as indicated by one of the reviewers, we should indicate that all our long units were extracted within the same fire season and no longer apart than 6 months. Therefore, it is very unlikely that the same area was burned twice in that period. The dating errors are certainly part of the accuracy (we dedicated a full section to them), but they should not be mixed with the spatial errors, which are mainly related to sensor or algorithm limitations. We further discuss this in the manuscript and indicate the reasons behind the lower reporting accuracy of our product (87% of detections within ± 10 days).

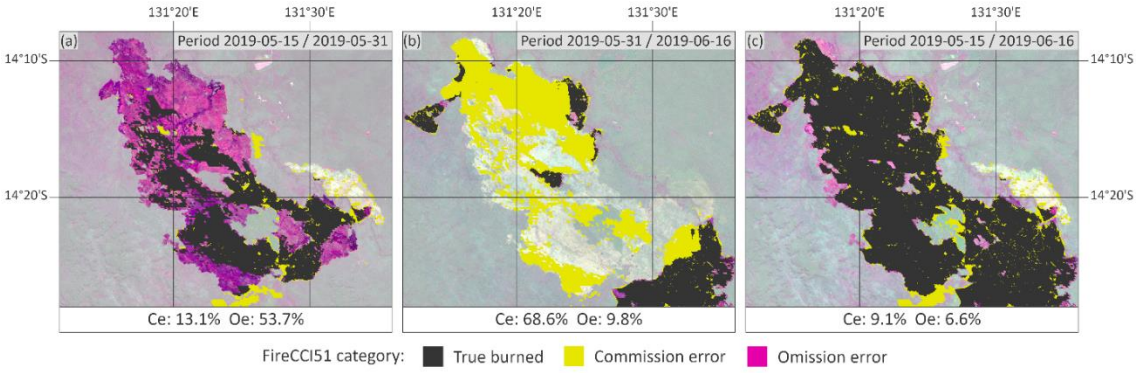

Figure 1. Figures (a), (b) and (c) show the post- and pre-fire Landsat 8 image difference (RGB: 7,5,4) at Landsat scene 105-070 (North Australia), for three different periods: (a) from 2019-05-15 to 2019-05-31 (16-day long), (b) from 2019-05-31 to 2019-06-16 (16-day long), and (c) from 2019-05-15 to 2019-06-16 (32-day long) that covers both (a) and (b). FireCCI51 pixels detected as burned for each period are overlaid on the images: black ones correspond to true burned, magenta to omission errors and yellow ones to false detections (commission errors). In (a) we see a large undetected burned surface that is dated by the product as burned in the second period (b), generating in this

second case a large commission error that is in fact a true burn. In contrast, when using a longer period (c), the influence of detecting burned-pixels days before or after a wildfire (temporal reporting accuracy) is minimized.

Other issues raised by the reviewers referred to the use of two versus one satellite, which creates differences in data obtained for 2017-18 (only Sentinel-3A) and 2019 (both Sentinel-3A and B). We have included a new analysis for 13 tiles, with additional validation data, testing the impact of the increased temporal frequency (more observations), concluding -as expected- in a higher accuracy. The global accuracy values of 2019 are lower than for the two other years, but this is also related to the global fire occurrence of both periods: what matters in this case is that C3SBA10 is more similar to FireCCI51 in 2019 than in 2017-2018, meaning that more similar results are obtained when the temporal resolution between sensors is more similar.

Finally, the manuscript was rewritten to better reflect the focus of the paper, figures were improved following reviewers' suggestions.

References

Franquesa, M., Lizundia-Loiola, J., Stehman, S. V., and Chuvieco, E.: Using long temporal reference units to assess the spatial accuracy of global satellite-derived burned area products, Remote Sensing of Environment, in review, 2021.

Hsu, N. C., Lee, J., Sayer, A. M., Kim, W., Bettenhausen, C., and Tsay, S. C.: VIIRS Deep Blue Aerosol Products Over Land: Extending the EOS Long-Term Aerosol Data Records, Journal of Geophysical Research: Atmospheres, 124, 4026-4053, https://doi.org/10.1029/2018JD029688, 2019.

Merchant, C. J., Block, T., Corlett, G. K., Embury, O., Mittaz, J. P. D., and Mollard, J. D. P.: Harmonization of Space-Borne Infra-Red Sensors Measuring Sea Surface Temperature, Remote Sensing, 12, 10.3390/rs12061048, 2020.

Sayer, A. M., Hsu, N. C., Lee, J., Carletta, N., Chen, S. H., and Smirnov, A.: Evaluation of NASA Deep Blue/SOAR aerosol retrieval algorithms applied to AVHRR measurements, Journal of Geophysical Research: Atmospheres, 122, 9945-9967, https://doi.org/10.1002/2017JD026934, 2017.

Sayer, A. M., Hsu, N. C., Lee, J., Bettenhausen, C., Kim, W. V., and Smirnov, A.: Satellite Ocean Aerosol Retrieval (SOAR) Algorithm Extension to S-NPP VIIRS as Part of the "Deep Blue" Aerosol Project, Journal of Geophysical Research: Atmospheres, 123, 380-400, https://doi.org/10.1002/2017JD027412, 2018.