Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-409-AC3, 2021

Biogeosciences Discussions

Interactive comment on "Quantifying the Importance of Antecedent Fuel-Related Vegetation Properties for Burnt Area using Random Forests" *by* Alexander Kuhn-Régnier et al.

Alexander Kuhn-Régnier et al.

ahk114@ic.ac.uk

Received and published: 8 April 2021

We thank the reviewer for their detailed reading of the text and correspondingly constructive feedback which will undoubtedly improve the quality of this work.

Referee comments are cited in *italics* and author's responses in normal font. Responses are separated by horizontal lines.

The authors use a machine learning approach (ML) to investigate the impact of bio-

physical and climatic variables on burned area over a five year period where there are available data on a global scale. FAPAR appears to be the most important predictor from their RF simulations. Variables were tested for their impact if they included the preceeding X number of months. This allowed for the investigation of the impact of lagged relationships, arguably very important for fire modelling.

I am concerned about the cross-validation strategy employed here (L137. By randomly choosing the validation dataset the potential for spatial autocorrelation issues arises. This is well known in the literature (see Roberts et al. 2017; Ploton et al. 2020; Kühn and Dormann, 2012; Meyer et al. 2019). Here is a snippet from the abstract to the Roberts article: 'Ecological data often show temporal, spatial, hierarchical (random effects), or phylogenetic structure. Modern statistical approaches are increasingly accounting for such dependencies. However, when performing cross-validation, these structures are regularly ignored, resulting in serious underestimation of predictive error. One cause for the poor performance of uncorrected (random) cross-validation, noted often by modellers, are dependence structures in the data that persist as dependence structures in model residuals, violating the assumption of independence. Even more concerning, because often overlooked, is that structured data also provides ample opportunity for overfitting with non-causal predictors.'. Because the authors devote considerable space to discussion of these predictors, I think this issue is worth consideration. The authors also argue [that] the gap in R2 of the training-validation simulations gives an idea of the generalizability of the model - but that breaks down if there are spatial autocorrelation issues. Also there is some spatial structure in their biases (Fig S2) that could be coming from this issue. I would suggest adopting other CV strategies as outlined in the papers I list above.

Comparing Fig. S2 (in the supplementary materials) to the normalised (by mean observations) differences (Fig. 2c in the manuscript), the spatial pattern changes if not disappears. The main spatial patterns apparent from Fig. S2 (in the supplementary materials) follow the magnitude of mean BA.

We will also adjust the number of bins in Fig. 2c (in the manuscript) in order to make the plot easier to read.

Further, the question concerning robustness verification is an important one, and we thank the reviewer for mentioning it. This will be addressed further below.

As I am not yet convinced by their CV strategy, which is important as it impacts the results quite heavily, I suggest major revisions as I assume it will take a bit of work to demonstrate that the chosen CV strategy doesn't give misleading results.

We have undertaken additional experiments to demonstrate the robustness of our model against overfitting. Excluding a variable number of grid cells around the testing data grid points (as in Ploton et al. (2020)) while keeping the number of training grid points constant yields the results shown in Fig. 1.

It can be seen that the performance of the model drops as a larger region around the test samples is excluded (with 40 pixels corresponding roughly to 1200 km at the equator). However, as opposed to the case study in Ploton et al. (2020), the performance remains relatively stable, not dropping below an R2 score of 0.4.

We have also undertaken an analysis of monthly data (as opposed to climatological averages) for the time period 11-2000âĂŤ12-2019 (230 months). A different burnt area dataset, MODIS MCD64CMQ BA, was used for this run. This experiment, the 15VEG_FAPAR_MON experiment, otherwise uses the same variables as the 15VEG_FAPAR experiment.

These results again show decreased R2 values, which is to be expected given the higher variance of this data:

- Using random cross-validation; Test R2: 0.45, OOB R2: 0.45
- Excluding the years 2009-2012 (including 2012); Test R2: 0.38, OOB R2: 0.45

- Excluding the years 2016-2019 (including 2019); Test R2: 0.41, OOB R2: 0.45

Despite the slightly reduced R2 values, the above experiments show that the model is able to robustly predict BA under spatial and temporal cross-validation scenarios. For the temporal cross-validation either period 2009-2012 or period 2016-2019 is left out for testing.

Minor comments:

- Line 78: I am not sure if I understand the DD calculation. If you had, say 5 days in one month below the precip threshold then 10 into the next. A brief precip event then the rest of the month below threshold. How does that rate? Would it be concatenated so the whole month is seen as being DD?

Concatenation is only carried out across month boundaries if the resulting dry-day period is contiguous. Wetting precipitation events always disrupt a dry-day period. We have added an example calculation to the manuscript to make this clearer:

"The dry-day period was defined as the longest contiguous period of ERA5 mean daily precipitation below 0.1 mm day^{-1} (wetting rainfall; Harris et al., 2014; Jolly et al., 2015) within each month. A period contiguous with the previous month's dry-day period was concatenated such that the sum of both (number of days) was used to determine the longest period. For example, consider a 30-day long month with a 10-day long dry-day period at the beginning of the month, followed by a wetting precipitation event on day 11, and then a dry-day period for the following 19 days. This month has a dry-day period of 19 days. However, if the previous month were to terminate in a 10-day long dry-day period, these 10 days would be added to the initial 10-day dry-day period of the current month, thereby making this combined dry-day period the longest."

- L80: Soil moisture was done how? Was it percent of saturation? relative to field capacity? Or some sort of index since it is later refered to as SWI.

It is the SWI index as provided by Copernicus, see mention in line 81 of the manuscript.

- L81 - The Kaplan and Lau reference is for the WGLC that is based on the WWLLN, not the WWLLN directly. It also says it provides the frequency of lightning flashes per unit area but doesn't specify that these are ground strikes. Can you please check that these are indeed cloud to ground and not total (cloud to ground + cloud to cloud)

That is true, WWLLN was used previously because it is the dataset that the WGLC is based on. The manuscript has been updated to use the WGLC dataset name. We have also added additional references regarding the type of lightning strokes detected by the network; the WWLLN network is unable to rule out detection of any cloud to cloud strikes but is most sensitive to cloud to ground strikes:

"We used the WGLC dataset (Kaplan and Lau, 2019) which provides counts of monthly lightning strikes. It is based on the World Wide Lightning Location Network (WWLLN) dataset, which mainly detects cloud-to-ground strikes (Rodger et al., 2004; Abarca et al., 2010), as opposed to LIS (Bürgesser, 2017)."

- Table 1: For the AGB datasets, was there any overlapping latitudes between the two datasets? If so, how was that dealt with?

Both datasets were first averaged from the original resolution to a 0.25° x 0.25° grid, then mosaicked to the global extent by taking the mean. This has been pointed out in the updated manuscript, along with the original resolutions of the individual datasets:

"Tree AGB was obtained by mosaicking AGB datasets for the tropics (Avitabile et al.,

2016, 1 km resolution) and for northern forests (Thurner et al., 2014, 0.01° resolution) using the mean after resampling each to a common spatial resolution of 0.25°."

- L105 - I am concerned about the gap-filling approach.So for SWI doesn't this mean that it would assume drought conditions? How often would you have this condition applied (outside of winter, L100)?

As shown in Fig. 2 for FAPAR, outside of winter the minimum value filling is virtually never applied. Additionally, it is mostly limited to northern latitudes in winter, as expected.

Regarding the filling of SWI, this should not have a big influence on the final results because we are not using antecedent values of SWI in any case. Since we don't expect fires during the winter, having (by necessity of gap filling) unphysical values of SWI in the winter should not affect results where relevant for our analysis. Other variables (e.g. low temperature during the winter) should be sufficient for the random forest model to ignore these SWI values.

- L110: I don't follow what was done here. What do those numbers mean? This is the smallest area of herb or crop in a pixel?

They are indeed the smallest observed area of HERB or CROP in a pixel. Missing land cover was previously filled using these values to prevent excessive data gaps.

However, this has now become obsolete due to renewed processing of the landcover dataset (to enable processing of a longer time period for specific runs). Consequently, this line has been removed from the manuscript.

- L125: I think X is referring to both the variables (FAPAR, etc.) and a single month,

i.e. I think you are saying variable X at month, t, has the seasonal cycle for variable X (called in the text, X 12M) subtracted from it? Please reconsider how this is formulated in the text. This doesn't work as written.

That is correct. We have improved the clarity of the manuscript by writing 'X 0M' instead of 'X' to reflect the fact that this refers to the variable 'X' in the current month.

-Table 2: To help choose which variables were useful for the ML algorithm (and which might only be contributing to overfitting), why did the authors not try applying one of the most standard techniques available such as recursive feature elimination with crossvalidation (available in scikit-learn, which the authors are already using (L132); Pedregosa et al. 2011)? These techniques could help get around arbitrary choices about the number of predictors, e.g. why 15 and not 10 or 20?

Part of our original intention was to evaluate how different importance measures affect the choice of variables. While RFECV is of course possible to carry out in principle, this is usually easily done with the Gini importance which is quite cheap (computationally) to calculate, as it only considers data already seen during training. Unfortunately, this also means that RFECV fails to account for overfitting, as it only considers the training data when calculating feature importance (Meyer et al., 2019). In contrast, the different approaches we utilise to calculate feature importance (aiming to calculate a more robust metric) are computationally demanding, making an approach such as RFECV infeasible.

While we agree that the choice of 15 predictors is somewhat arbitrary, this is heuristically based on the slope of the feature importance plots (see Fig. 3). Looking at these feature importances, where the importance change is 'flat' (by inspection) after 15 variables, no additional information was being conveyed. Therefore, we decided to use this as our threshold.

Given this choice of 15 variables (and some constraints as discussed in the manuscript), we also carry out an iterative approach where we investigate different sets of variables to determine which combination of vegetation variables yields the best results. This was done to balance the computationally intensive nature of the CV calculations with the ability to answer the question "Which vegetation proxy is the most important?".

- Why does Fig S2/2/3/7/etc. have straight line cutoffs? Top of Mexico, Australia, E of Iran. Plotting error? Real problem? I don't recall this being mentioned in the main text but it is in all figures. Fig 3-Also for v. low BA, what about grey instead of black? Would be easier to see the larger BA values...

The 'block-shapes' are caused by our choice of AGB dataset, which we have mentioned in all relevant figure captions in the updated manuscript. We have also changed the way BA is plotted to more correctly highlight areas with missing data as shown in Fig. 4.

We have added grey shading to indicate regions with fire data availability, but where one or more of the other datasets is not available. Light grey indicates regions where mean BA is 0, with dark grey representing regions with nonzero mean BA.

Note that in addition to the shown changes in this figure, the colour scheme for plot c will also be revised to make the differences more apparent, and the number of divisions adjusted in order to make the plots clearer.

- L 300- Are those the right refs? Both papers use Causality Analysis and not regression techniques as mentioned here. What aspects of those papers touches on inclusion of extra predictors and overfitting in a ML approach?

The causality analyses in those papers are also based on regression techniques so that the same logic applies. In fact, causality methods are often used for predictions in

the same way as random forests (e.g. Kretschmer, Runge, and Coumou (2017)). The problem is always: there is a curse of dimensionality in high dimensional regression problems, and this can lead to overfitting if the dimensions are not controlled in some way. Our interpretation of random forests here is also causal by considering feature importances as a measure of driver importance behind the earth system process.

Both papers state, for example, that the explanative power of the machine learning approach they use (Causal Networks) decreases as additional, irrelevant variables are included. For example, Nowack et al. (2020) mentions "low detection power [. . .] if too many variables are used" and Runge et al. (2019) explains "Ideally, we want to condition only on the few relevant variables that actually explain a relationship".

We expect these observations to apply to our methodology, because inclusion of irrelevant variables may increase the likelihood of fitting to noise in the data, which, although likely somewhat mitigated by the bootstrapping employed by the random forest algorithm and cross-validation, may still present an issue.

- p 12 has a lot of 'discussion' in the results section. Either have a 'results and discussion' section or keep them separate.

As recommended by the reviewer, we have adapted the manuscript to have a combined 'Results and Discussion' section.

- Fig 6 - can you not use sci notation for the numbers? It makes it easier to read. I am not sure if this warrants main text inclusion as it is very simple with almost no interesting structure. I would put this in supplement and move S2 or S3 into the main.

These figures will be adapted to no longer make use of scientific notation in the tick labels.

While results shown in Fig. 6 (in the manuscript) may not be revolutionary or complex, they do illustrate one of the key findings – the ability to empirically determine and visualise intuitive interactions between variables. We therefore believe this figure still warrants inclusion in the main text.

- L310 - what about masking for areas with longer fire return intervals to see if it [then] pops out as more important?

Unfortunately, even considering the 20-year long MODIS record, we are strongly limited by the data available to us. Predictability in boreal ecosystems is expected to remain very limited because the return times are many times longer than the time series, so there is a very large stochastic component. Therefore, we don't currently expect this to be a viable analysis.

References

Abarca, Sergio F., Kristen L. Corbosiero, and Thomas J. Galarneau. 2010. 'An Evaluation of the Worldwide Lightning Location Network (WWLLN) Using the National Lightning Detection Network (NLDN) as Ground Truth'. Journal of Geophysical Research 115 (D18): D18206. https://doi.org/10.1029/2009JD013411.

Avitabile, Valerio, Martin Herold, Gerard B. M. Heuvelink, Simon L. Lewis, Oliver L. Phillips, Gregory P. Asner, John Armston, et al. 2016. 'An Integrated Pan-Tropical Biomass Map Using Multiple Reference Datasets'. Global Change Biology 22 (4): 1406–20. https://doi.org/10.1111/gcb.13139.

Bürgesser, Rodrigo E. 2017. ‘Assessment of the World Wide Lightning Location Network (WWLLN) Detection Efficiency by Comparison to the Lightning Imaging Sensor (LIS): WWLLN Detection Efficiency Relative to LIS’. Quarterly Journal of the Royal Meteorological Society 143 (708): 2809–17. https://doi.org/10.1002/qj.3129.

Harris, I., P.D. Jones, T.J. Osborn, and D.H. Lister. 2014. ‘Updated HighâĂŘresolution Grids of Monthly Climatic Observations – the CRU TS3.10 Dataset’. International Journal of Climatology 34 (3): 623–42. https://doi.org/10.1002/joc.3711.

Jolly, W. Matt, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, and David M. J. S. Bowman. 2015. ‘Climate-Induced Variations in Global Wildfire Danger from 1979 to 2013’. Nature Communications 6 (1): 7537. https://doi.org/10.1038/ncomms8537.

Kaplan, Jed O., and Hong-Kiu Lau. 2019. ‘The WGLC Global Gridded Monthly Lightning Stroke Density and Climatology’. PANGAEA. https://doi.org/10.1594/PANGAEA.904253.

Kretschmer, Marlene, Jakob Runge, and Dim Coumou. 2017. ‘Early Prediction of Extreme Stratospheric Polar Vortex States Based on Causal Precursors’. Geophysical Research Letters 44 (16): 8592–8600. https://doi.org/10.1002/2017GL074696.

Kühn, Ingolf, and Carsten F. Dormann. 2012. ‘Less than Eight (and a Half) Misconceptions of Spatial Analysis’. Journal of Biogeography 39 (5): 995–98. https://doi.org/10.1111/j.1365-2699.2012.02707.x.

Meyer, Hanna, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. 2019. ‘Importance of Spatial Predictor Variable Selection in Machine Learning Applications – Moving from Data Reproduction to Spatial Prediction’. Ecological Modelling 411 (November): 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815.

Nowack, Peer, Jakob Runge, Veronika Eyring, and Joanna D. Haigh. 2020. ‘Causal Networks for Climate Model Evaluation and Constrained Projections’. Nature Communications 11 (1): 1415. https://doi.org/10.1038/s41467-020-15195-y.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. ‘Scikit-Learn: Machine Learning in Python’. Journal of Machine Learning Research 12: 2825–30.

Ploton, Pierre, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, et al. 2020. ‘Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models’. Nature Communications 11 (1): 4540. https://doi.org/10.1038/s41467-020-18321-y.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta GuilleraâĂŘArroita, Severin Hauenstein, et al. 2017. ‘Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure’. Ecography 40 (8): 913–29. https://doi.org/10.1111/ecog.02881.

Rodger, C. J., J. B. Brundell, R. L. Dowden, and N. R. Thomson. 2004. ‘Location Accuracy of Long Distance VLF Lightning Locationnetwork’. Annales Geophysicae 22 (3): 747–58. https://doi.org/10.5194/angeo-22-747-2004.

Runge, Jakob, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. ‘Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets’. Science Advances 5 (11): eaau4996. https://doi.org/10.1126/sciadv.aau4996.

Thurner, Martin, Christian Beer, Maurizio Santoro, Nuno Carvalhais, Thomas Wutzler, Dmitry Schepaschenko, Anatoly Shvidenko, et al. 2014. ‘Carbon Stock and Density of Northern Boreal and Temperate Forests’. Global Ecology and Biogeography 23 (3): 297–310. https://doi.org/10.1111/geb.12125.

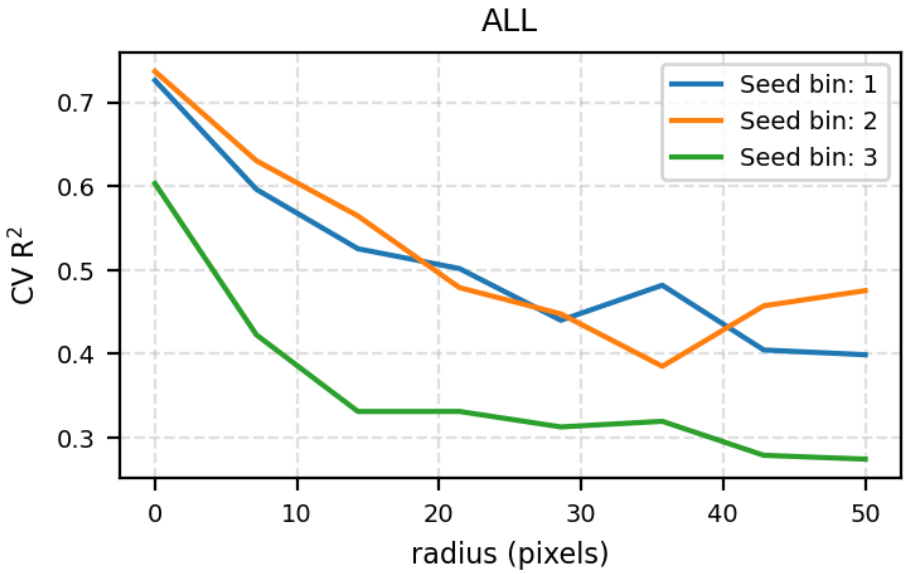

Fig. 1. R2 scores for burnt area (BA) prediction on test samples for different exclusion radii around the test samples. Each line represents the R2 score for around 333 individual test samples.

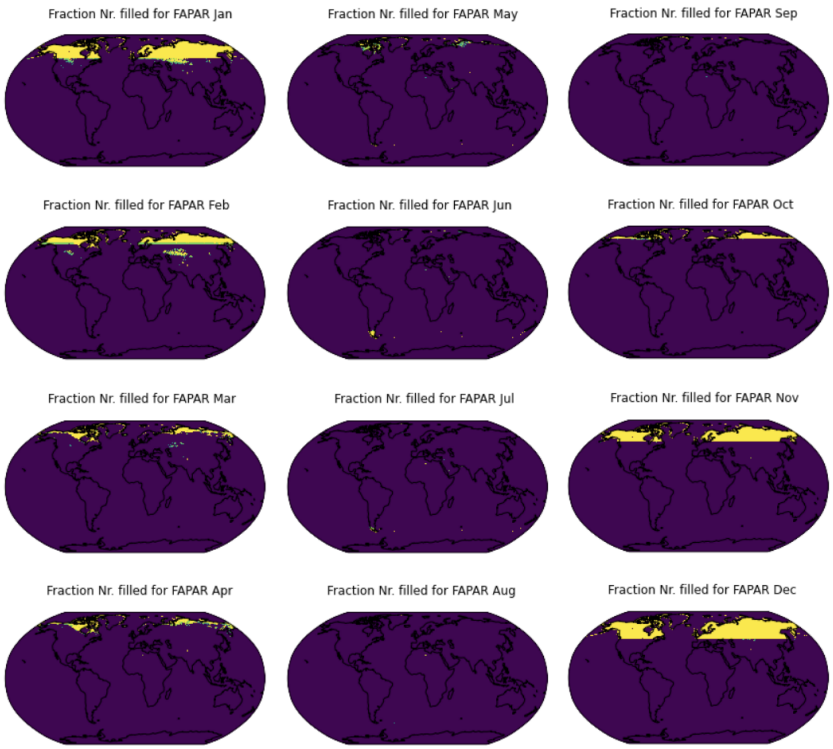

Fig. 2. The proportion of filled samples for FAPAR, with yellow indicating that all occurrences of a given month at a given location were filled and purple indicating no filling was done.

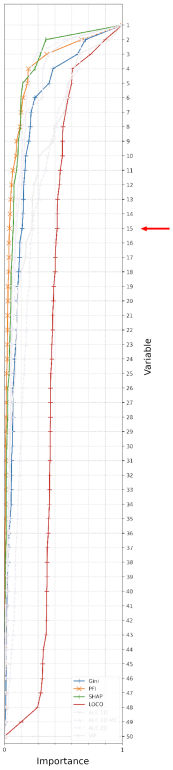

Fig. 3. Sorted variable importance metrics (Gini, PFI, SHAP, and LOCO) for the ALL model, with the highest variable importance according to each metric at the top. Note alternative metrics are greyed out.

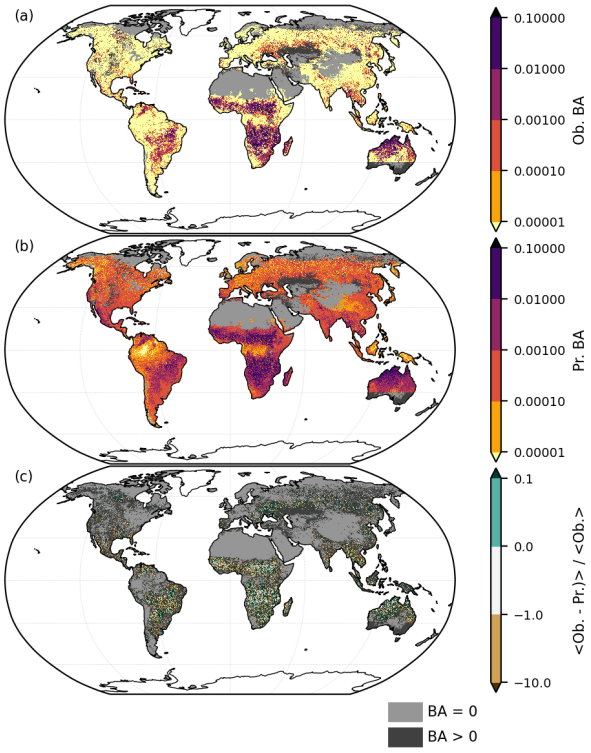

Fig. 4. (a) Average observed (Ob.) BA derived from the GFED4 BA dataset. (b) Out-of-sample predictions (Pr.) by the ALL model. (c) Relative prediction error of the ALL model.