Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-885-RC1, 2019

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Predicting wildfire burned area in South Central US using integrated machine learning techniques" *by* Sing-Chun Wang and Yuxuan Wang

Anonymous Referee #2

Received and published: 27 November 2019

Summary: This manuscript described a method for estimating burned area in the southern central region of the United States using three machine learning methods applied serially, with training derived from an existing dataset. The results show some skill in modeling total burned area over large areas. The work is focused mainly on the role of climatic variables in estimating burned area totals. While the methods in this paper might be of interest to the broader community, the manuscript is not well written (the structure is difficult to follow and it requires significant language editing throughout), the results cannot be reproduced because there is not enough information about the input variable processing, and the significance and limitations of the study are not explained well. Again, this method could prove to be useful to the broader community,

but the manuscript needs significant work and for that reason I recommend rejecting this paper.

General Comments: While I think the methods presented in the manuscript have potential to produce useful results, the manuscript needs to be improved in order to create a more logical flow of information, better describe the data used, illustrate the output, provide a more complete literature review, and provide details about the usefulness and limitations of this study. Furthermore, it requires editing beyond the scope of scientific peer-review.

The goals of this study are unclear – is the goal to predict wildfires in the future based on weather conditions, to support climate projections, or to simply estimate the amount of burned area? A related critique is that the structure of the paper makes it difficult for the reader to follow, there are effectively two methods sections with the results of the first set of methods in the middle. Additionally, the authors never present a figure showing the modeled burned area, which should be the main output of this work and really needs to be emphasized in the main body of the manuscript.

The authors have not considered a large body of wildfire research regarding satellite observations-driven modeling which is relevant to this work in the background research. Similar studies involving the effects of climate on total burned area should be noted by the authors, including Andela et al., 2017 and Zubkova et al., 2019. Additionally, the methods section refers to aspects of the data which are not described until a later section. This organization is difficult for the reader to follow, and the description of the data used is insufficient, in part because the sources of the input data are not provided. The data preprocessing methods are unclear as well – how is a discrete thematic variable like land cover type represented at 0.5-degree resolution? How are the translations between quantiles and area being made, given that the area of the grid cell varies with latitude? Also, the authors say the model predicts burned area at 50 km spatial resolution (with no indication of the map projection used), this is not the same as 0.5 degrees and this discrepancy needs to be resolved. There are also questions

about the fire data used to train the model – are prescribed fires included in the data (by definition, these are not wildfires in most cases)? Is there an estimate of the number of fires which are omitted? Given the quantile-based approach, what happens if there is a fire or amount of burning which is greater than any in the training dataset (i.e. it would fall out of the range of the training data unless there was a training cell with 100% burned area)? Is the length of the training dataset long enough to capture all variability in fire activity as it relates to climatic conditions? Why were remote sensing-derived datasets not considered?

An important aspect of fire regimes which was not adequately considered in the manuscript is the role of human activity in the fire regime, especially in the United States where humans play an active role in the fire regime through suppression, ignitions, fuel load management, and landscape fragmentation in addition to being the source of ignition of approximately 85% of fires (according to the US Forest Service). These effects vary as function of not only population density, but sociopolitical norms which can vary from state to state. Recent papers such as the Andela et al. 2017 paper claim human activity is the major control on fire activity, and as such it cannot be ignored in a study region where the fire regime is likely human-driven. While the datasets describing human activity are certainly far from perfect, it is not possible to describe fire activity in a human-driven fire regime without considering human influences.

Finally, there needs to be more effort in describing the expected impact of the work and the limitations of the method. For example, the abstract mentions that the work can be used to assess fire management strategies but provides no details on how or why. The quality of the input data is not discussed, which will propagate errors through the model, as well as the serial structure of the integrated model itself. At present, the manuscript is too focused on the machine learning exercise rather than on the scientific value of the work.

Specific Comments:

L75: Why were other months of the year excluded?

L81: “Uneven data” is used throughout the paper but is not defined. Does this refer to unevenness spatially, temporally, or both?

L92-95: Given that the model compares the output to the quantile ranges, is it capable of estimating an amount of burning greater than has been observed in the training data?

L155: Is there any concern about propagation of error through the model? What is the benefit of running three models in serial rather than one model alone or several models in an ensemble?

L164: Is there an estimate of the number of fires missed? Small fires constitute most of the fires by number, even though they add up to relatively little burned area (e.g. Malamud, Millington, and Perry 2005). It is noted that the dataset omits most small fires occurring on private land – these are not generally wildfires and such fires should be omitted anyways if the study is about wildfires.

L194-195: I don’t think the climatic variables can be considered as fixed, especially since the assumption in later parts of the paper surround climate change scenarios which means their values do vary through time.

L209: Is any consideration given to preventing overfitting due the correlation between variables? For example, ecoregions and landcover types are likely to be related to one another.

L232: Please clarify the phrase “horizontal scale of around 700 x 700 km2” – the use of horizontal scale implies a one-dimensional unit (length) which does not match the unit specified. Also, as a suggestion, 700 x 700 km2 seems ambiguous and could be more clearly represented as “700 km x 700 km” or “490,000 km2”

L252: SUS is never defined

L268-270: One could argue that the model is in fact “hardwired” (editorially, the term is jargon and should be replaced) to the geographical features of the study domain – geography deals with the human components of space and time as well as the physical components. The tendency of the human population to ignite or suppress fires as a result relationship to sociopolitical factors (like local regulations) will influence the fire regime in ways which will not be captured by climatic variables and will change from location to location.

L286: Why were 14 variables chosen? This seems like an arbitrary cutoff, especially given the large number of variables which went into the model.

L306: The fuel-related variables are among the least important presented in Figure 5 – how can the conclusion be drawn that fuel abundance is what determines the amount of burned area?

Table 1: The resolution of the data is presented, but it’s not clear how the data are being re-gridded to the working resolution - if the fire analysis is being done at 0.5 degree and the climate data is at 32 km resolution then there are < 4 cells per burned area data point and the way which those $\sim$4 cells are represented has significant consequences.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-885, 2019.