Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., author comment AC2
https://doi.org/10.5194/nhess-2021-245-AC2, 2021

Reply on RC2

Tadas Nikonovas et al.

Author comment on "ProbFire: a probabilistic fire early warning system for Indonesia" by Tadas Nikonovas et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-245-AC2, 2021

The authors would like to thank reviewer #2 for their positive view of the manuscript and for constructive feedback and suggestions. Please find below a point by point response to the comments received.

Reviewer comments received are italicised while our responses are in normal font. Line numbers in our responses refer to those in the revised manuscript.

MODIS imagery is used as a basis for fire inventory. I am wondering whether this choice might introduce a bias towards large fire events, consequently leading to an underestimation of small fires? This is mentioned around line 118, but might be worth a short comment in the discussion section.

This is a valid point as omission of small fires and consequent underestimations of fire activity is a well known problem associated with MODIS fire datasets. However, we believe that it is of little importance in the context of this study, which aims to predict elevated fire activity rather than quantify fire effects such as burned area or fire affected area. Importantly, as our results show, the number of grid cells with low monthly active fire counts (< 11) does not exhibit large interannual variability, and hence it is not of much importance for seasonal forecasting. Prediction of rare high activity fire events is the main interest of early warning systems and as a result, omission of small fires in MODIS record is less important. We have added a sentence to state this (lines 122-125):

" In any case, omission of small fires in the product is less of a problem for early warning systems which aim to alert of unusually high activity fire events, rather than quantify fire effects such as fire-affected area."

l. 144: how is the threshold motivated? Is this just because 10 is a nice number, or is there evidence that counts > 10 are particularly dangerous or may indicate particularly

dangerous conditions?

The reviewer is right to point that the active fires > 10 threshold was selected based on numerical properties. However, this was done not because of number 10 per se, but because it divides the whole dataset of monthly active fire counts at 25km resolution into 95% (active fires < 11) vs 5% (active fires > 10) parts. Coincidently, monthly active fire count > 10 threshold also represents top 25% (upper quartile) of all grid cells where active fires were detected within a given month (active fires > 0). While we have no evidence suggesting that the threshold represents particularly dangerous fire conditions on the ground, but we believe it is nonetheless a useful indicator of "elevated" fire risk as it indicates the top quartile of all fire affected grid cells.

The threshold does also represent a "sweet spot" between the increasing dataset imbalance and potential usefulness of the forecasts. During the initial model testing phase we noted that at this level of dataset imbalance the model was still generating generally reliable and calibrated probabilities. While prediction of even more rare events (increasing the threshold) would have been potentially more useful, however this resulted in loss in forecast sharpness (very little to no high probability predictions) and hence reduced the potential economic value of the forecasts.

We have reworded section 2.1.1 to make the selection criteria clear (lines 129-150)

Section 2.3: There are some indicators available through Copernicus Global Land Service (https://land.copernicus.eu/global/themes/vegetation) which might be of interest for future work. Most of them are based on Sentinel-3/OLCI or PROBA-V and are available from 2014 onwards.

Thank you for this suggestion, the Copernicus Global Land Service is indeed a promising source of information for future fire prediction studies.

Section 3.1: It is unclear to me whether the hyperparameters were tuned or simply set to the values reported in the manuscript? The text does not provide any indication of hyperparameter tuning (e.g. via cross validation), a quick glance into the code implies that values were simply determined beforehand. This should be stated more clearly.

We did perform a cross validation using grid search to determine optimum number of hidden layer neurons, alpha parameter and solver for the mlp classifier. The Brier score was the benchmark for parameter selection. This is now stated in the manuscript: "The model architecture and optimal parameter setup were determined performing grid search cross-validation and evaluating the model's performance on validation data."

Section 3.1: Minor technial nitpicking: I think the split described her refers to `train' and `test' sets, respectively. The `validation' set is usually a subset of the training dataset used for hyperparameter tuning, before testing the tuned and validated model on the unseen `test' data.

Thank you for pointing out this discrepancy. Changed the wording accordingly.

Section 3.6 / Formulas (3) and (4): I think it is more common to use the terms "true positive" (TP) instead of "hits"; "false negative" (FN) instead of "misses" and "false positive" (FP) instead of "falsealarms". "pod" could be "sensitivity" or "recall".

We agree with the suggestion regarding "true positive", "false negative" and "false positive". Changed accordingly. However, the authors feel that "probability of detection" is a more intuitive term for non-specialist audiences and it makes sense to use it here, in particular given that we are dealing with rare event prediction. While "sensitivity" and "recall" are more widely used in machine learning circles, we suggest to keep "probability of detection" in the manuscript considering that the readership of the journal (geosciences, environmental sciences, policy and broader public).

I assume that the data set is somewhat imbalanced - i.e., there are more non-event pixels than fire pixels. Was this accounted for (in terms of model formulation or in terms of performance metrics)?

This is a very important observation and the dataset indeed is imbalanced as we discuss in Section 2.1.1. While the level of imbalance is not extreme (95% vs 5% for active fires > 10 case) it is nonetheless considerable. However, the dataset is relatively large (449280 samples overall) and the minority class is still well represented by over 21000 grid cells. We did take special care in selecting forecast evaluation metrics which are appropriate for probabilistic predictions regardless of dataset imbalance. Note that imbalance can cause problems when dealing with classification methods that turn continuous probability distribution to dichotomous predictions by setting arbitrary class cut-off thresholds and optimize models using improper scoring rules (accuracy is one example). In this study we used reliability diagrams and the Brier score (a proper score) which by definition optimizes for "true" expected event occurrence probabilities and work regardless of class imbalance (when comparing performance of different models/model parameters for the same sample of events).

We did not subject the data to any special treatment (such as oversampling the minority class) nor did we see a need to adapt the model formulation to account for imbalance. Artificial balancing of the class sizes prior to training or application of different class weights to the loss function during training would only lead to artificially inflated minority class probabilities, and consequently would have a negative impact on the performance metrics employed in this study.

Finally, I would like to acknowledge the provision of the ProbFire source code via GitHub and the comprehensive data avilability statement.

Thank you for the appreciation.