Sustainable Forestry (2023) Volume 6 Issue 1
doi: 10.24294/sf.v6i1.2490

ORIGINAL RESEARCH ARTICLE

Forecasting wildfire hazard across northwestern south America

Andrea Markos[1,*], William Matt Jolly[2], Ernesto Alvarado[3], Harry Podschwit[3,4], Sebastian Barreto[1], Catherine Toban[1], Blanca Ponce[1], Vannia Aliaga-Nestares[5], Diego Rodriguez-Zimmermann[5]

[1] United States Forest Service, International Programs, Thomas Circle NW, Washington, DC 20005, United States. E-mail: anmarkos@gmail.com

[2] United States Forest Service, Rocky Mountain Research Station, Missoula Fire Sciences Laboratory, W. Broadway Street, Missoula, MT 59808, United States.

[3] School of Environmental and Forest Sciences, University of Washington, West Stevens Way NE, Seattle, DC 98195, United States.

[4] Oak Ridge Institute for Science and Education, Bethel Valley Road, Oak Ridge, TN 37830, United States.

[5] Servicio Nacional de Meteorología e Hidrología (SENAMHI) del Perú, Jr. Cahuide 785, Jesús María 02002, Lima, Peru.

ABSTRACT

Fire hazard is often mapped as a static conditional probability of fire characteristics' occurrence. We developed a dynamic product for operational risk management to forecast the probability of occurrence of fire radiative power in the locally possible near-maximum fire intensity range. We applied standard machine learning techniques to remotely sensed data. We used a block maxima approach to sample the most extreme fire radiative power (FRP) MODIS retrievals in free-burning fuels for each fire season between 2001 and 2020 and associated weather, fuel, and topography features in north-western south America. We used the random forest algorithm for both classification and regression, implementing the backward stepwise repression procedure. We solved the classification problem predicting the probability of occurrence of near-maximum wildfire intensity with 75% recall out-of-sample in ten annual test sets running time series cross validation, and 77% recall and 85% ROC-AUC out-of-sample in a twenty-fold cross-validation to gauge a realistic expectation of model performance in production. We solved the regression problem predicting FRP with 86% r^2 in-sample, but out-of-sample performance was unsatisfactory. Our model predicts well fatal and near-fatal incidents reported in Peru and Colombia out-of-sample in mountainous areas and unimodal fire regimes, the signal decays in bimodal fire regimes.

Keywords: Wildfire Hazard; Google Earth Engine; Machine Learning; Operational Risk Analysis; Out-of-sample Validation

ARTICLE INFO

Received: 25 February 2023
Accepted: 30 March 2023
Available online: 9 April 2023

1. Background

The aim of this study is to establish a grounding for a dynamic operational product that estimates fire hazard expressed as the probability of occurrence of "near-maximum fire intensity locally possible"[1], using fire-danger weather indices and including topography, wind/topography interactions and fuel canopy remotely sensed data. The probability estimation updates as new data becomes available and is intended to aid fire managers' decision-making process by reducing the gap between coarse descriptors of the environment and predictive outputs. Fire intensity (fire radiative power, FRP) is regarded as the main controller of fire spread through either positive or negative feedbacks[2], in a relationship mediated by wind speed. At higher fire intensity, the heading fire rate of spread may be accelerating or decelerating, depending on complex interactions between wind, characteristics of the fire environment and the fire itself[2]. We may not know which one is the case when the high fire radiative power measurement has almost one km resolution and the granule is daily, we do know that in both cases the locally possible near-maximum FRP is hazardous. We used finer-scale descriptors of the fire environment to reduce the scale from coarse weather products to match the resolution of the science-ready FRP MODIS-14A1.061 products and modelled fire hazard at that spatial and temporal scale, i.e., daily about one km resolution.

This study was conceived after two important milestones in the fire risk literature[1,3]. In these studies, the fire-danger weather index energy release component for fuel model G was input to a univariate logistic regression model to determine the probability of large fires to occur, training the model on historical fire data for different fire-size thresholds[1,3]. FlamMap5 was then used to simulate fire spread with an adjustment for 'non-forest' fire[1], simplifying the fuels' distribution, assumed to be model G for CONUS. Fuel model G contains fuel load in all size classes, i.e., conventionally defined as the time required by the fuel to reach a balance with the moisture in the surrounding environment, e.g., 1 h through 1,000 h timelag[1,3]. Scott *et al.*[1] define wildfire hazard as a geospatial output, input to the analysis of exposure and effects in risk analysis. Scott *et al.*[1] considered desirable to "assess the near-maximum wildfire behavior (e.g., fireline intensity) possible at each pixel on a landscape". We interpreted that probability estimation of daily near-maximum expected fire output in a dynamic sense for operational risk assessment, contingent upon daily fuel and weather conditions. Wildfire hazard is the product of the probability of burning and the expected wildfire intensity given that a burn occurs[1] both are a function of topography and the fast-changing fire environment: fuel moisture and weather. In hazard assessment the concern is how likely it is to observe the near-maximum locally possible fire intensity for free-burning vegetation, we thus sampled fire intensity maxima recorded during twenty-one fire seasons to build and combine two models: a burn probability estimation (biased toward the highest observed values) and a regression model intended to predict the intensity of those locally possible extreme values. The output of the model is intended to assess wildfire hazard on a daily basis.

The near-maximum potential wildfire intensity for a point, stand, or landscape has been typically assessed as fireline intensity, rate of spread or flame length of the heading fire under 80th, 90th, and 97th percentile of the energy release component (ERC) conditions[3], or just the "near-maximum" conditions 97th percentile[1]. We built upon[1] narrower focus. In addition to ERC, the 97th percentile 1 min average wind speed at 6 m occurring during the typical burning period of the typical fire season (locally defined in terms of months of the year and hours of the day) has been applied in the upslope direction on all pixels regardless of aspect. Scott *et al.*[1] also used the 97th percentile dead fuel moisture contents for all size classes of dead surface fuel.

In terms of fire-danger metrics, flame length is derived from fireline intensity, but most the uncertain conditions for fireline estimation, hardly met in controlled observational settings (see the study of Finney *et al.*[2] for an in-depth discussion on this point), are out of reach with remote sensing retrievals. We focused instead on fire radiative power retrievals as a proxy for fireline intensity to develop a predictive model intended for operational applications. An intensely burning wildfire detected at 927 m resolution is a (relatively) slow-onset phenomenon, it may or it may not be related to fast spreading, behave erratically and catch firefighters off-guard[4–10], but often it does not necessarily affect large areas, if the final burn scar perimeter is considered[7,11] unlike fireline intensity, flame length or fire rate of spread, the extent of the final perimeter of the burn scar by itself is not a fire danger metric[12,13]. Wilson[14] identified common denominators of fire behavior linking firefighter's fatalities and entrapments, signaling issues of great concern into the present day[7–9,12,14–16]. Most fatalities and injuries occur:

1) On relatively small fires or deceptively quiet sectors of large fires.

2) In relatively light fuels, such as grass, herbs, and light brush.
3) When fire responds to topographic conditions and runs uphill.
4) When there is an unexpected shift in wind direction or in wind speed.

According to Colombian official statistics, the sixteen reports of deadly fire incident between 2003 and 2021 affected on average forty-five hectares; if forty-six reports that include fire injuries are considered, the average fire size reaches 139 hectares (source: IDEAM, n.d.). The size of the dependent variable we modeled is about eighty-six hectares (MODIS: 927 m nominal scale), which is a considerable size for a wildfire to anyone who happened to get trapped in it. In Peru, between 2003 and 2020, 98% fatalities and injuries occurred in light fuels, most of which above 2,500 m on rough land (source: INDECI, n.d.). Points one to three in Wilson's list represent features relevant to fire hazard that we have encoded as strata into the sampling procedure or otherwise inform the study criteria:

1) We considered burning pixels individually, not necessarily belonging to the head fire of a large wildfire.
2) Stratifying by fuel-type (reclassified as "forest" vs. "no-forest", following the study of Scott *et al.*[1] to make the model more robust to fuel type classification uncertainty.
3) Sampling terrain features systematically to establish a relationship to extreme FRP retrievals.

Unfortunately, the coarse spatial and temporal resolution of MODIS thermal anomalies makes it very infrequent to identify active fire on 927 m pixels with an average percent slope >10%. We sampled slope as one more predictor feature but did not use it for stratification. Shifts in wind direction or speed might be encoded as events using sub-hourly weather data like NOAA/SWM real-time mesoscale analysis (available only for CONUS). Modeling unexpected shift in wind direction or in wind speed is a pending task. Established tools allow to estimate wind/topography interactions and their effect on fire behavior (WindNinja), rate of spread, time of fire arrival and even expected fire perimeter after n-days (FlamMap, Behave Plus, FSPro). These existing models can be used to obtain probability of high-intensity fire. These tools cannot be used in most of the world where vital inputs are not available: 1) fuels data including layers such as fire behavior fuel models, canopy height, base canopy height, and canopy bulk ratio, 2) decades of high-quality weather observations, and 3) decades of catalogued fire occurrence.

We intend to forecast probabilities in real time for operational applications, conditional upon recent multispectral vegetation indices retrievals, topography, wind/topography interactions, and the most relevant fire danger indices forecasts, to be recalculated in synch with the weather forecast data stream for situational awareness and operational risk analysis. The outputs generated by our model are not static probability maps of fire characteristics occurrence simulated over tens of thousands of years[1,3]. This study makes a first attempt at integrating elements of fire danger to a revisited wildfire risk model, to answer at any time the question: "Provided that a wildfire occurs, what is the probability of observing fire radiative power in the locally possible near-maximum fire intensity range?"

2. Methods

2.1 Research design for statistical power

Most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class[17]. The importance of this consideration could not be overstated: a classification model, e.g., a logistic regression, may solve a binary classification problem with an overall an accuracy of 99%. This result may be misleading in the case that the binary output that we intend to forecast equals zero 99% of the time, and one 1% of the time. In this case the prior would be 99%, thus a classifier with no skills could classify the outcome correctly 99% of the times by mere chance. This is the kind of problems inherent to imbalanced datasets. One may find out that despite the 99% overall accuracy, such model misclassifies 100% of the minority class, when the outcome equals one. This result could not be worse if the perfectly misclassified output is a fire danger measure such as “near-maximum fire radiative power”. A high rate of false negatives would mean that no alert is issued when fire potential is extremely high, consequences might be deadly. This issue is akin to that of a wrong medical diagnosis: the patient is positive to a deadly, contagious disease but misclassified as negative. This type of error is called false-negative, and it can be much more expensive than the false-positive error[18].

Logistic regression is the first algorithm developed to estimate the probability of occurrence, it was conceived to solve prediction/classification problems in randomized trials[19]: “Success or failure (outcome one or zero) is recorded at each trial and it is required to test the null hypothesis that successes occur randomly with probability 0.5 against the alternative that there is a trend in the success rate.” To estimate the probability of a binary output the dataset needs to be balanced, so that the prior implies a correct 50% probability to belong to either one or the other class. The machine learning literature provides numerous techniques to correct the problems of imbalanced datasets[17] usually by randomly under-sampling the majority class to match 1:1 the minority class. We implemented random under-sampling early on, during the data generation process.

To derive a probability using a classification algorithm, the highest FRP retrievals need to be matched to burnable pixels that did not burn within the time window of that sampling run, to ascertain at least partially why one burned while the other did not. In the parlance of experimental studies these would be called “counterfactuals”, short for “counter-to-fact conditional”, what would have been true, had certain facts been different. In causal explanation the model is requested to estimate the conditional probability that a particular event in time was the cause of a particular outcome, cast as a counterfactual question: had A been false, would B still have happened?[18] The outcome of concern for us is “the near-maximum wildfire intensity possible at each pixel on a landscape”. Such outcome is conditional upon local weather, topography, and fuels, to explain why this pixel exhibited such extreme behavior (historical maxima are not absolute or theoretical, only empirical, remote measurements) while another equally burnable pixel did not.

We needed to create a balanced dataset. In doing so, we needed to minimize the risk of confounding factors influencing the results. We needed to ensure that our results have high probability of detecting an effect, when in fact one exists. A model with high statistical power has a low probability of making a Type II error, i.e., a low false negative rate. A false negative classification error equals to not expecting wildfire hazard when due. Such unforeseen surprises must be minimized. Additionally, no more than the right amount of data is appropriate to ensure swift algorithm deployment in real time using a pre-trained model. We thus needed to yield the maximum statistical power with just the right amount of data to generate reliable forecasts on unseen data, out-of-sample.

To solve at once all our optimization goals, we embedded statistical matching techniques in the data generation process to address the main problems of observational data analysis that lead to poor generalization: imbalance in the distribution of the confounders, and model dependence in the statistical estimation of causal effects[20]. Conducting randomized trials like those that led to the first applications of logistic regression[19] in a wildland setting to study and forecast the highest FRP observed in free-burning wildfires, and matching those to a valid counterfactual is made possible only within a statistically defined quasi-experimental design, safely applied to remotely sensed FRP data. We thus applied established techniques of statistical matching to achieve the same results sought after in the randomized control trials literature to improve causal inference in observational studies and reduce model dependence in the statistical estimation of causal effects[21] by: 1) constructing the best possible comparison group based on observed characteristics[22], pre-screened on the basis of their predictive power; 2) reducing imbalance in the distribution of the pre-treatment confounders between the treated and control groups. The relevant math is the same in the machine learning and randomized trials literature, but the latter has been around for longer and is better understood.

To reduce selection bias, it is very important to match the outcome of interest to a valid comparison group based on observed baseline characteristics, reducing the unobservable differences[22]. We thus imposed these strict conditions: 1) the counterfactual pixel belongs to the same fire regime and predominant fuel type; 2) it is currently vegetated; 3) it can burn, as proven by independent historical record; and 4) it was not burning by the time the sampling takes place, nor it did shortly before; hence it safely qualified as a "zero" at the time of sampling. We identified one main fire regime in the study area, corresponding to the months comprised between July and November during the time of interest (**Figure 1**). We identified six strata combining west, north, and south, and distinguishing forest vs. no-forest (**Figure 2**). Thus, we sampled the highest FRP values observed in each stratum between July and November during twenty fire seasons. Readers are encouraged to use the first companion web app to explore relevant data (see conclusions). **Figure 1** shows the seasonality of the main fire regime in the study area as a sum of higher quality FRP retrievals between 2001 and 2020. The fire energy output distribution is bimodal overall, but we focused on the period between July and November in the main unimodal fire regimes.

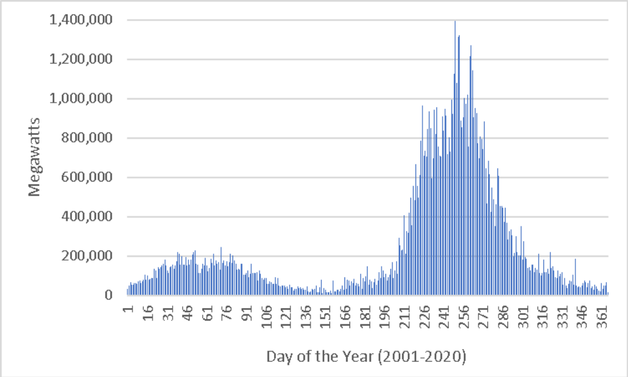

Figure 1. Sum of highest quality MODIS-14A1.061 fire radiative power retrievals at 6 km scale (2001–2020).

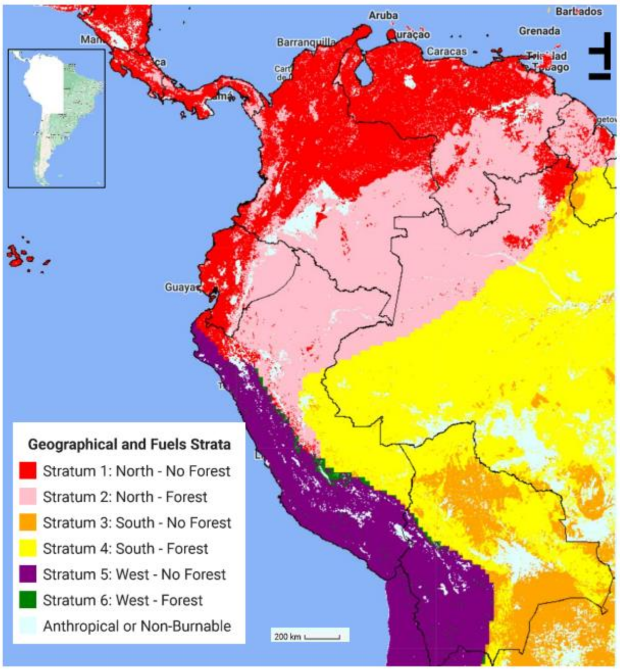
Aruba
Barbados
Grenada
Barranquilla
uraçao
Caracas
Trinidad
Tobago
Guaya
Geographical and Fuels Strata
🟥 Stratum 1: North - No Forest
🟥 Stratum 2: North - Forest
🟧 Stratum 3: South - No Forest
🟨 Stratum 4: South - Forest
🟪 Stratum 5: West - No Forest
🟩 Stratum 6: West - Forest
⬜ Anthropical or Non-Burnable
200 km

Figure 2. Study area, geographical and fuels strata.

2.2 Confounding factors and stratification

2.2.1 Fuel strata

Free-burning mask

In the context of predictive services, our model will calculate a probability of free-burning in any fuel, including crops and pastures; but in order to develop it, we excluded detections of fires that most likely have been started under controlled circumstances, like slash burning and other intentional uses of fire. Controlled burning in the region is almost exclusively tied to grazeland management, straw or slash and burn practices, so we excluded from the training data any pixel that might have ever been classified as urban, agriculture, water, or ice, at any time between 2000 and 2020. Prescribed burning in the study area is practiced on a small scale, and typically in low-risk days; thus it is extremely unlikely that any prescribed burning detects slipped in our training set that selectively sampled only the thirty-four highest FRP retrievals of each fire season.

With the aim to sample only free-burning natural vegetation, we obtained twenty burnable layers intended to mask false-positives as well as agricultural uses of fire. We integrated all LULC collections using a one-out-all-out method[23]: none of the filtered pixels of interest for this study was ever classified as urban, agricultural, water, ice or any other anthropical land use combining the most recent (2001–2020) MapBiomas collections PanAmazonia v3, Brasil v6, and Chaco v2[24–26], the Copernicus global land cover collection[27], and the MODIS land cover product MCD12 v006[28]. For greater confidence, we increased a 1,000 m buffer (larger than the typical MODIS pixel size) around any 30 m or greater pixel that has ever contained a single 30 m pixel classified as anthropical use from the mapbiomas collections. The yearly masks and distance layers thus generated were resampled to match the target variable resolution.

Non-burning counterfactuals

The algorithm used to measure fire radiative power can correctly classify active fire which might be as little as 100 m^2 in a 927 m pixel[29], while the fuel available for burning in the entire pixel might never be completely consumed by fire. A single pixel can be correctly classified as actively “burning” tens of times in a single year. In fact, burned area products such as MCD64A1 and FireCCI often classify correctly as “burned” the same pixel multiple times within a year and in subsequent years. Fire rotations in free-burning fuels can be surprisingly high, given that the most frequent cause of ignition in the study area is anthropical. An exhaustive analysis of this issue is beyond the purpose of this paper, but the data are freely available for confirmation.

To ensure that the counterfactual pixels have not been burning at the time of their inclusion in our sample or the previous month, we ruled out any burning pixel obtained from any fire product available in the Google earth engine (GEE) platform, see **Table 1**. Additionally, we allowed only pixels that burned at least once between 2001 and 2020, according to the same datasets used to retrieve the dependent variable.

Table 1. Collections queried to reduce the sampling error for the counterfactual (non-burned pixels).

Dataset (GEE code)	Provider(s)	Reference(s)
ESA/CCI/FireCCI/5-1	ESA	[30]
MODIS/006/MCD64A1; MODIS/061/MCD64A1	NASA LP DAAC–USGS EROS Center	[31,32]
FIRMS	NASA/LANCE /EOSDIS	[33]
NOAA/GOES/16/FDCF	NOAA	[34,35]

Forest vs. no-forest fuels strata

We made our model robust to fuel type misclassification and relied instead on continuous variables such as multispectral vegetation indices. An issue of concern is that the accuracy of fuel type remote classification is typically low, even more so for countries that do not have the resources to develop finer products. For example, the LANDFIRE remap prototype presents an overall classification accuracy of 52% in the LULC layer that serves as the basis for fuel type classification[36]. Other inputs used for fire modelling such as the LANDFIRE products canopy base height (CBH) and canopy bulk density (CBD) also have relatively low map validation accuracy, such as r^2 0.09 for CBH and r^2 0.34 for CBD, averaged across most map zones in CONUS[37]. We thus considered prudent to reduce the reliance upon, and any potential bias due to fuel type misclassification and rely instead on multispectral vegetation indices that provide an approximation to canopy characteristics with reduced latency.

To ensure that forest fires are equally represented along with the most frequent savannah fires, we adopted a simplified binary fuel type classification: "forest" vs. "no-forest". Once a LULC product is resampled to 927 m to match the independent variable most details are lost, thus it is only the predominant fuel type that is reclassified. This strategy to deal with uncertainty is not different from modeling ERC for fuel type G[1,3] and then introducing coarse adjustments for forest[1]. We decided to use satellite retrievals to approximate recent fuel characteristics, leveraging freely available products that cover the entire time of interest such as MODIS-TERRA collections (232 m and 463 m resolution; see section on multispectral vegetation indices).

Consistency of classification across three years is a commonly used criterion under land cover classification uncertainty[38] that we adopted to classify forest vs. no-forest as predominant. We used three years sliding window to dichotomize the LULC classification obtained from the FAO-LCCS3 nomenclature available in the MCD12Q1 collection (423 m, see **Table 2** for more details). The stratification delivered as expected, making fuel-type almost irrelevant to probability estimation, thus debiasing the model (see results).

2.3 The target variable

Paraphrasing Scott *et al.*[1], wildfire hazard is the potential for a wildfire to cause harm to people or property. The primary factor that determines wildfire hazard is wildfire intensity, usually some measure related to the rate of energy released by a fire. The greater the wildfire intensity, the greater the potential for harm. Other factors that affect wildfire intensity include fuel, weather, and topography. Wildfire hazard at a given location is thus quantified as the product of two characteristics:

- Burn probability: the likelihood that a fire will occur at that location.
- Conditional wildfire intensity given that a fire does occur: the distribution of wildfire intensity, such as flame length or fireline intensity, at that location.

Wildfire likelihood and intensity are combined into a single measure of wildfire hazard, used to identify areas where there is the potential for high wildfire hazard and prioritize management opportunities to reduce and manage wildfire risk, protecting people and property.

In this study, we took an empirical approach to estimate wildfire hazard based on high quality remote measurements of fire radiative power and estimated the probability of such observed extreme fire outputs contingent upon topography, and daily fuel and weather conditions. Wildfire hazard is a dynamic phenomenon unevenly distributed across the landscape and in time. Factors such as fuel moisture, wind speed, and temperature affect wildfire hazard and change on a daily (even sub-daily) basis.

A critical metric of wildfire hazard estimation is fireline intensity (FLI). FLI is operationally defined as "the rate of heat release per unit length of flaming fire front, calculated as the product of heat content, fuel consumption during flaming front passage, and rate of spread"[1]. The main drawback to FLI is that it cannot be physically measured, only estimated[39]. We queried the science ready MODIS collections 14A1 v061 which provide about 20 years of daily remote measurements of fire radiative power (FRP) in megawatts[32]. Unfortunately, attempts at deriving sub-pixel fractional area of wildfires from MODIS retrievals delivered poor results[40,41]. Ignoring the sub-pixel burning area is a mayor hindrance to the estimation of fireline intensity. Consequently, MODIS-retrieved FRP will be modelled in its stead as it represents the highest quality proxy for fire intensity available at scale.

In addition to the filters described above to include only free-burning fire in natural vegetation, we applied filters to exclude false fire detections (water, bright soil, iron roofs, bare rocks, etc.). One such filters is the level of confidence attributed to the published product that fire is correctly detected and there are no concerns with regards to the quality of fire radiative power estimation[33]. The MODIS fire products 14A1 v061 are stage-three validated, meaning that uncertainties in the product are well quantified through solid ground-truthing and other suitable reference data. MODIS 14A1 v061 data are ready for use in scientific publications.

We admitted only the highest confidence data according to the data provider, corresponding to the “FireMask” attribute equal nine, into the sampling routine, a function passed through the entire image collection that sorts and selects the thirty-four most intensely burning hotspots within each stratum, each fire season. This approach to extreme value analysis is called block maxima; it focuses on identifying the most extreme values within specific time periods or seasons[42]. This allows for the analysis of the rare but significant fire events that have the potential to cause considerable damage or impact.

2.4 Predictor features

Table 2 shows the complete list of image collections queried to develop the model. Most of these predictors can be updated in real time with no cost. The model is simple and can run in any suitable programming environment. The main difference with real-time deployment consists in using weather forecasts instead of reanalysis data. The walkthrough between ERA 5 land and the weather forecasts available in real time exceeded the purpose of this paper.

Table 2. Image collections used to develop the model.

ERA5-land hourly	ERA5-land hourly–ECMWF climate reanalysis (11,132 m): apps.ecmwf.int/datasets/licences/copernicus/
Weather fire danger indices	Calculated from ERA5-land hourly 11,132 m: www.wfas.net/data/SAR/
Wind speed and wind/topography matching index	ERA5-land hourly–ECMWF climate reanalysis (11,132 m): apps.ecmwf.int/datasets/licences/copernicus/ The matching index was calculated in the GEE platform
Terrain products	NASA SRTM DEM (30 m): doi.org/10.1029/2005RG000183
Multispectral vegetation indices	8-day composite MOD09A1.061 (232 m) https://doi.org/10.5067/MODIS/MOD09Q1.061 8-day composite MOD09A1.061 (463 m) https://doi.org/10.5067/MODIS/MOD09A1.061
Distances from agricultural areas, urban areas, and inland water bodies	Calculated from mosaicked MapBiomas (30 m) and MODIS-derived MCD12Q1 v006 LULC collections (463 m): MapBiomas: chaco.mapbiomas.org/; pampa.mapbiomas.org/; mapbiomas.org/; amazonia.mapbiomas.org/; MCD12Q1 v006: lpdaac.usgs.gov/products/mcd12q1v006/

2.4.1 Weather fire danger indices

Twenty years (2001–2020) of daily fire danger indices were derived from the hourly ERA5-land reanalysis at 11,132 m resolution[43]. Weather data were obtained using the CDS-API and hourly data for the study region were composited to daily extremes. These daily weather data were then used to calculate the fire weather index (FWI)[44], energy release component (ERC) and burning index (BI) following[45], as well as the keetch-byram drought index (KBDI)[46]. Finally, we derived the spread component (SC)[47] from BI and ERC to account for wind speed and changes in fine fuels moisture.

2.4.2 Wind/topography interactions: The angular difference model

The features we derived from the angular difference of wind direction and aspect is an adaptation from the study of Jolly *et al.*[48]. The aim is to capture the interactions of wind and terrain. Topographical aspect is the direction that slope is facing, expressed in degrees, and increasing clockwise from the north. Wind direction is expressed in the same way. When wind interacts with a slope in perfect alignment, it is impact on fireline intensity, fire spread rate and flame length is maximum, pushing fire uphill. On the lee side of the orographic formation, the effect of wind is at its minimum because wind speed is decreased, and its direction is downslope. When aspect and wind direction perfectly align, the angular difference is 0 and when they are out of alignment, it can increase to 180. This feature is encoded as a percentage: angular difference is divided by 180 to convert it to a value that can vary between 100% for perfect alignment and 0% to identify the lee side. Maximum wind speed and matching index have been sampled for the day when maximum FRP values have been observed filtering the diurnal hours between 08:00 and 20:00 in the UTC-5 time zone, because more relevant to fire activity (in line with the study of Scott *et al.*[1] and Finney *et al.*[3]). We resampled with the bicubic option the hourly ERA5-land reanalysis at 11,132 m resolution and made it interact with NASA SRTM DEM, resampled to match the independent variable.

2.4.3 Multispectral vegetation indices

We computed or retrieved multispectral vegetation indices (VIs) adopting the published formulas (**Table 3**), querying MODIS-TERRA 8-day composites (see **Table 2**). Additionally, we evaluated the predictive power of the ratio of two popular vegetation indices, both related to fuel moisture content and highly correlated to each other (−97% in our dataset, these VIs carry almost the same information): the normalized burn ratio (NBR) vs. the normalized difference moisture index (NDMI). NDVI, NBR and NDMI were calculated retrieving the corresponding multispectral bands from the 8-day composites MOD09A1.061 (232 m) and MOD09A1.061 (463 m) with three weeks latency to simulate real-time deployment of the model.

Table 3. Vegetation indices and references.

Vegetation index	Formula	Reference
Normalized difference vegetation index (NDVI)	(RED-NIR)/(RED + NIR)	[49]
Normalized burn ratio (NBR)	(SWIR2-NIR)/(SWIR2 + NIR)	[50]
Normalized difference moisture index (NDMI)	(NIR-SWIR1)/(NIR + SWIR1)	[51]

2.4.4 Distances from agricultural areas, urban areas, and inland water bodies

It is of interest to compute features that characterize sampled pixels in relation to agricultural and urban areas, as frequent sources of ignition on the landscape. Distance from water bodies has also been encoded as a predictor feature, relevant to human-ecosystem interactions and vegetation moisture. Rivers represent a navigable network constituting the equivalent of roads in remote areas, mostly in the amazon basin lowlands. Proximity to water bodies secures water access for all human activities.

We generated twenty yearly layers to estimate the impact of distance to urban, agricultural and water pixels on fire danger. We combined the land cover products with a 30 m resolution such as MapBiomas PanAmazonia v3 (2001–2020), Brasil v6 (2001–2019) and Chaco v2 (2001–2019), and with a 463 m resolution the MODIS collections (2001–2019) such as MCD12Q1 v006 using a one-out-all-out method. These distance features were computed for each year in the series using the "fast distance transform" in GEE. It returns the distance, as determined by the specified distance metric (defaults to squared Euclidean distance), to the nearest non-zero valued pixel in the input. The output contains values for all pixels within the given neighborhood size (256), regardless of the input's mask. The scale of the pixel was set equal to the independent variable.

2.5 Assessing the random forest algorithm's out-of-sample performance

Consistently with literature[52–54] and with the results obtained from earlier analyses, random forest performed dependably to forecast probability of occurrence of the most intensely burning hotspots each fire season, in each stratum. We conducted feature and model selection independently to solve the classification and the regression problem: 1) probability of occurrence; and 2) retrieved fire radiative power. In both cases, the first step consisted of ranking features in terms of importance on the first training subset and ruling out multicollinearity by applying the backward stepwise procedure, starting with all available predictors, and recursively eliminating the least performing feature in any pair that correlated near or higher than 60%. We maintained all features able to pass the multicollinearity filter and that contributed to the predictability of wildfire hazard. We proceeded similarly when implementing the time-series cross validation, also called walk-forward validation[55] and the twenty-fold cross validation, to get the most out of the available dataset. In both cases, we implemented recursive feature elimination and model regularization using the scikit-learn library[56].

Performance metrics usually improve with trees number, at least up to a certain point, past which accuracy may even decline with too many trees[57]. However, the increase in computation time is not negligible. Too many trees may lead to over-fit, under-fit or “out of memory” errors. We decided to keep the number of trees low in comparison to popular modelling software default values: Minitab’s default parameter is two hundred and five hundred in R. We settled for three hundred random trees.

We assessed the performance of the classification model with threshold metrics, setting the threshold at a standard 50%, to ensure that both false positive and false negative rates are balanced and as low as possible. False negative rates are of most interest as not issuing an alert when appropriate may cost human lives. The true negative rate is also termed specificity and the metric used to assess specificity is called recall, of special interest to our case-study as recall’s emphasis is on false negatives. Precision and recall are combined into a single score that seeks to balance both concerns, called the F-score:

Accuracy = Correct Predictions/Total Predictions

Precision = TruePositive/(TruePositive + FalsePositive)

Recall = TruePositive/(TruePositive + FalseNegative)

F-score = 2 × Precision × Recall/(Precision + Recall)

We additionally used a popular ranking metric: the receiver operating characteristic area under the curve (ROC-AUC). A ROC curve is a diagnostic plot analyzing the performance of a model, plotting the false positive rate against the true positive rate. The true positive rate is the recall (see formula above), alternatively called sensitivity. The false positive rate is calculated as:

FalsePositiveRate = FalsePositive/(FalsePositive + TrueNegative)

The area under the ROC curve provides a single score to summarize the plot that can be used to compare models. We adopted a standard 50% threshold, equivalent to the prior of our quasi-experimental approach. A no skill classifier will have a score of 0.5, whereas a perfect classifier will have a score of 1.0. The ROC AUC can be optimistic under a severe class imbalance, especially when the number of examples in the minority class is small[17]. That risk is minimized with our balanced dataset.

Cross-entropy is perhaps the most common metric used to evaluate predicted probabilities for binary classification, measured as LogLoss or the negative log likelihood[17]. LogLoss summarizes the average difference between two probability distributions: observed vs. predicted. A perfect classifier has a LogLoss of 0.0, with worse values being positive up to infinity. For a binary classification dataset where the observed values are y and the predicted values are $\hat{y}$, LogLoss can be calculated as follows:

LogLoss $= -((1 - y) \times \log(1 - \hat{y}) + y \times \log(\hat{y}))$

We conducted power calculations using the online calculator of the statistics department of the university of British Columbia. Power calculations perform these operations: 1) estimate the average probability estimated by the model for both observed classes; 2) determine whether the two outcomes are different subtracting the averages; 3) test the null hypothesis against the alternative hypothesis and determine type II error probability (power);

and 4) establish how large must the sample be to ensure that observed effect is due to relevant covariates, rather than to lack of precision in estimates. Based on early versions of this study, we conducted power calculations to determine the sample size for desirable statistical power and significance[58], setting both close to 100%. We thus proceeded to sort and filter the thirty-four highest FRP retrievals for each fire season, in each stratum of interest.

We run all the performance tests for probability estimation (probability reliability, accuracy metrics, AUC-ROC and power calculations) on the model output obtained through time series cross-validation, to simulate the performance of the model as is walks-forward when deployed in real-time and gauge a realistic expectation about the performance of this advisory system in deployment. We included ex-post power calculations to ensure that the β parameter is as low as expected using out-of-sample estimated probabilities.

3. Results

3.1 Solving the classification problem: Regularizing and evaluating the model

To assess the added value of the approach presented here, we established a baseline using all the available fire-danger indices to determine the probability of observing the outcome of concern. The first run included the five indices and obtained a recall of 65%, a respectable edge over the 50% unskilled predictor. The only problem with this result is that it was obtained using highly inter-correlated fire danger indices (**Figure 3**). Multi-collinearity reduces statistical power by making coefficients unstable. The only compatible non-collinear (non-redundant) fire danger indices were ERC and SC, which have the lowest correlation to each other, 0.55 in our dataset, **Figure 3** ranks the predictors in terms of importance, normalized to the most important one. The spread component is last because the output selected is mostly plume-dominated wildland fire:

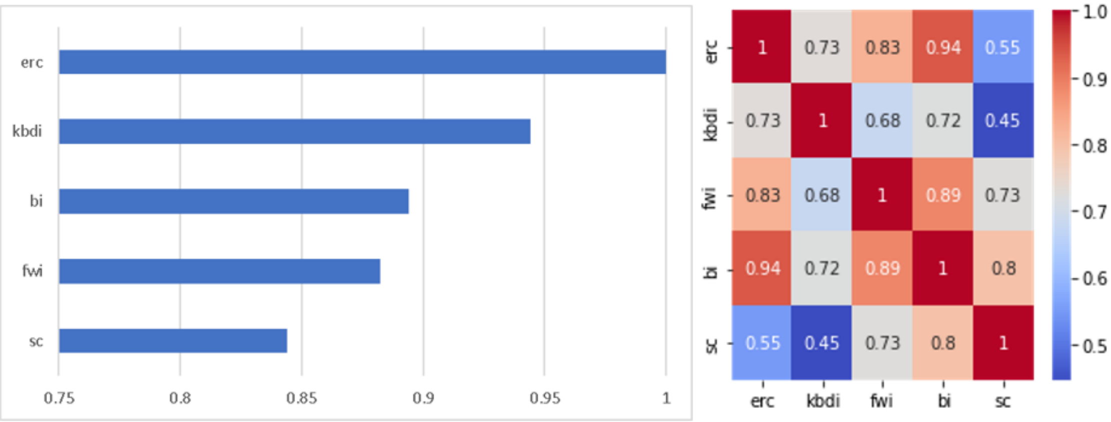

Figure 3. Relative importance of the baseline predictors and their correlations.

We then used ERC and SC to model the probability of the outcome of concern. The most relevant predictor, aided by the only non-collinear covariate available. We averaged the out-of-sample results of the twenty-fold cross-validation obtaining a LogLoss: 0.65, ROC-AUC: 64%, Recall: 61%, Precision: 60%, Accuracy: 60%, and F-score: 60%. The probabilities of a type I or type II error were close to zero. So, ERC and SC make a good combination as they achieve the most reliable modelling performance using only coarse resolution fire danger indices.

After establishing this baseline, we proceeded to incorporate all the available predictor features, iterating the same procedure to select features and regularize the model: we introduced all the available independent variables and then eliminated one by one the least important ones in each pair of variables that exhibited a correlation near or above 0.6. We averaged the out-of-sample results of the twenty-fold cross-validation obtaining a LogLoss: 0.50, ROC-AUC: 85%, Recall: 77%, Precision: 77%, Accuracy: 77%, and F-score: 77%. The probabilities of a type I or type II error were close to zero, but all the out-of-sample performance metrics improved significantly.

3.2 Probability calibration

To assess the reliability of the random forest's probability output, this is routinely compared to the observed probabilities as frequency of occurrence within the binned target variable[17]. Charting the probability calibration curve using the out-of-sample predicted probability obtained from the twenty-fold cross-validation returned a s-shaped distribution typical of the random forest probability output (**Figure 4**), indicating that probability calibration may be in order prior to model deployment. We thus implemented probability calibration through the standard procedure using Platt-scaling[17], but this actually worsened LogLoss. We then tried to use random forest for this task and that slightly improved the LogLoss score from 0.496 to 0.494. **Figure 4** shows the probability calibration curve using random forest instead of logistic regression, with the result of reducing the typical s-shape. Henceforth we will show results for the calibrated probability estimation only.

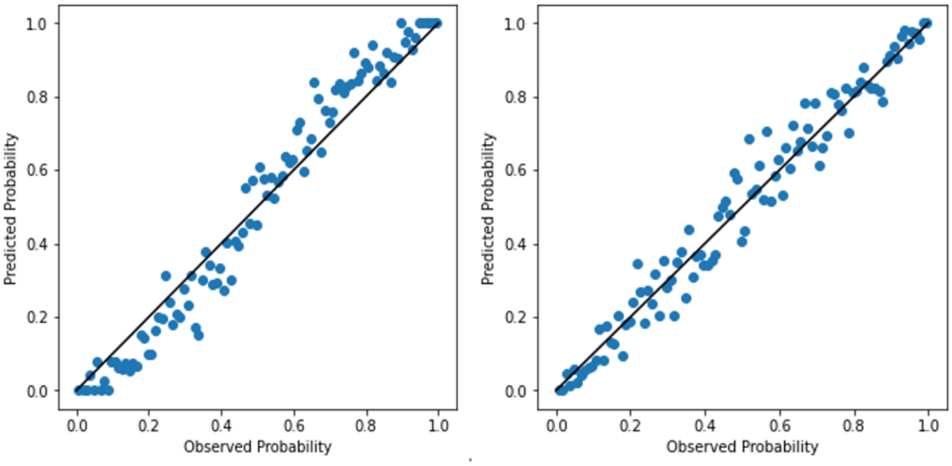

Figure 4. Probability calibration curve, uncalibrated (left), calibrated (right).

3.3 Model interpretation: Features importance

After model regularization and probability calibration, we proceeded to interpret the model obtained. **Figure 5** ranks the predictors in terms of importance, normalized to the most important one to facilitate interpretation. The three most important predictor features are related to fuels moisture being NBR, ERC and the ratio of NBR to NDMI. The next two most important features are longitude and latitude, commonly used to downscale coarse weather variables such as temperature and relative humidity. Multispectral vegetation indices can have a much higher resolution than weather forecast used to calculate fire-danger indices, estimating canopy characteristics such as foliar moisture. The normalized burn ratio (NBR) proved slightly more relevant to the outcome of concern than the normalized difference moisture index (NDMI). They highly correlate to each other ($r = -0.97$), so only one could be used in the model, while the non-collinear information carried by the NDMI could be rescued using the NBR/NDMI ratio instead. NBR, ERC and the ratio of NBR to NDMI are the most important predictors in the model. The three of them relate to fuel moisture ignoring fuel type, just characterizing the canopy. Interestingly, the previous 12-month peak EVI (proxy for biomass accumulation) did not improve model's performance metrics and correlated with NBR (−0.54), so it was removed from the predictors' array. The 8-day VIs products and the ERC usually change slowly (ERC requires seven days of data for its calculation). This slow change pace is compensated for by wind-derived features and the

SC, also to account for daily changes in fuel moistures. **Figure 5** shows the relative importance of each predictor in relation to the most important fuel-related feature (NBR).

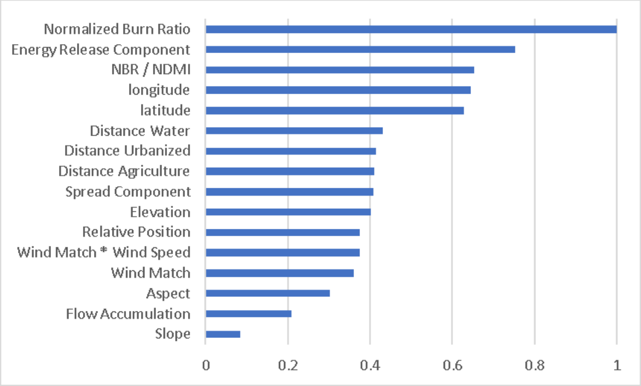

Figure 5. Relative importance or predictor features.

The next two most important features are latitude and longitude, while elevation follows closely, unsurprisingly, these three features (*x*, *y*, and *z* coordinates) ranked consistently within the top predictors in all modelling explorations conducted for this study. They are among the most important independent variables also in regressive models used to statistically downscale temperature, relative humidity, and other weather variables[59]. These predictors, along with the four least important predictors in our model (terrain products) have their legitimate place in the model because they contribute the same type of information required to downscale temperature and relative humidity, or ERC, and relate them to FRP outcomes.

Proximity to human settlements and agricultural activities influences fire hazard. These predictors change slowly but significantly on a year-to-year basis, especially the ever-expanding agricultural frontier. Wind speed and wind/topography interactions are more important than terrain products other than elevation. These predictors change by the hour and influence fire danger outcomes and fine fuels moisture, we aggregated them by the day. All predictors included in the regularized model correlate to each other within an acceptable range: −0.51 for latitude and longitude, 0.55 for SC and ERC, and 0.38 for ERC and NBR. The highest correlation tolerated is 0.65 between slope and elevation, the decision to include slope followed thorough testing how its inclusion impacted positively on all metrics, following the same procedure used for every doubtful correlated features.

3.4 Out-of-sample model evaluation

We first evaluated the performance out-of-sample recursively, simulating real time deployment of the model, training the model on historical data (e.g., 2001–2010), and evaluating it on new data (e.g., 2011). We compared the performance of a rolling window, which grows as new historical data become available, as opposed to a sliding window that included only the most recent ten years. A rolling window performed slightly better than a sliding window to “walk-forward” the model. **Table 4** shows walk forward recall, which overall averaged 75% out-of-sample in this time-series cross-validation. We achieved comparable results using a twenty-fold cross validation, assessing all the performance metrics of interest (**Table 4**).

Table 4. Twenty-fold all metrics cross-validation and walk-forward accuracy validation using a rolling window.

Random validation sets	ROC-AUC	Precision	F-score	Accuracy	Recall 1	Recall 2 (W-F)
Validation set 1	76%	66%	69%	68%	73%	-
Validation set 2	78%	67%	73%	70%	79%	
Validation set 3	80%	69%	74%	72%	80%	
Validation set 4	80%	70%	73%	71%	76%	
Validation set 5	80%	68%	75%	72%	84%	
Validation set 6	90%	81%	83%	83%	85%	
Validation set 7	89%	83%	79%	80%	75%	

Table 4. (*Continued*).

Random validation sets	ROC-AUC	Precision	F-score	Accuracy	Recall 1	Recall 2 (W-F)

Validation set 8	87%	80%	81%	81%	81%		
Validation set 9	92%	86%	86%	86%	85%		
Validation set 10	86%	82%	74%	76%	68%		
Validation set 11	83%	72%	75%	74%	79%	2011	77%
Validation set 12	80%	68%	75%	72%	83%	2012	78%
Validation set 13	82%	72%	75%	74%	78%	2013	75%
Validation set 14	85%	77%	81%	80%	85%	2014	79%
Validation set 15	84%	74%	78%	77%	82%	2015	75%
Validation set 16	85%	84%	76%	78%	69%	2016	71%
Validation set 17	89%	87%	77%	80%	69%	2017	77%
Validation set 18	90%	89%	79%	81%	71%	2018	76%
Validation set 19	87%	88%	76%	79%	67%	2019	70%
Validation set 20	89%	89%	76%	79%	66%	2020	74%
Out-of-sample	85%	78%	77%	77%	77%	-	75%

3.4.1 Solving the regression problem

We solved the regression problem predicting fire radiative power with 86% r^2 in sample, but the performance of the regression model out-of-sample was unsatisfactory, as shown in **Figure 6**.

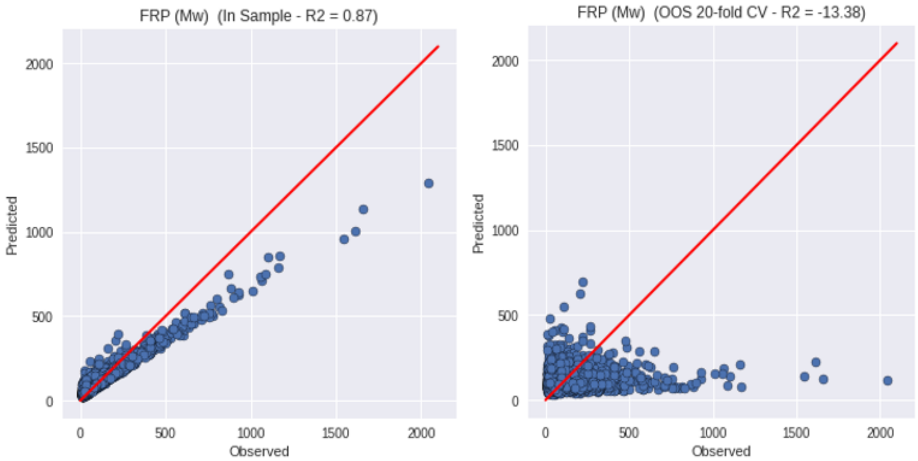

Figure 6. Modeled FRP in sample (left) and out-of-sample (right).

3.4.2 Independent validation set

We used an independent fire report dataset compiling official information from Peru (INDECI, n.d.) and Colombia (IDEAM, n.d.) with 50 and 58 reports of fatalities and injuries caused by wildfires respectively (2003–2020). When issuing an early warning detailed information about fire hazard, estimation needs to be aggregated according to a meaningful administrative unit, being municipal governments usually the most directly responsible to respond to fire incidents. We thus aggregated the spatial data at about 56 km (four global forecast service pixels, as this is the main candidate weather forecast for real time deployment) around the official coordinates of the fire incidents that reported fatalities and injuries. For the sake of model validation, we settled for a sufficiently coarse resolution to be relevant to most municipal administrative units.

We compared our model’s false negative rate to those generated with a largely validated methodology for fire danger classification[8,9]. We aggregated the four fire danger indices based on the adjective classification percentile thresholds over the available 21 years climatology: low 0–60th, moderate 60th, high 80th, very high 90th, and extreme

97th. Our model's false negative rate was 30% failed forecasts at 50% probability hazard threshold. The false negative rates of BI, ERC, KBDI and FWI at the "high danger" (80th) or higher threshold, was 34%, 36%, 44% and 47%, respectively. **Table 5** shows these results.

Table 5. Comparison of the hability to alert prior to a fatal of near-fatal wildfire incident in Colombia and Peru (2003–2020).

Adjective classification	**BI**		**ERC**		**KBDI**		**FWI**	
	N/108	**Correct**	**N/108**	**Correct**	**N/108**	**Correct**	**N/108**	**Correct**
Low	9	22%	10	30%	14	43%	21	48%
Moderate	28	79%	29	72%	33	73%	30	77%
High	29	79%	32	81%	23	74%	15	87%
Very high	26	69%	24	71%	22	73%	26	77%
Extreme	16	69%	13	69%	16	81%	16	63%

We report the percentage of incidents correctly predicted with 50% probability or more by our fire-hazard model within each adjective fire-danger class. The most important improvement can be noticed in the "low" and "moderate" fire danger categories, when established methods would not forecast fire danger.

Error analysis revealed that our fire hazard forecasts worked almost perfectly in unimodal fire regimes, but false negative rates increased in bimodal fire regimes, observed in bimodal precipitation regimes in parts of Colombia. This slice of data corresponding to markedly bimodal fire regimes in Colombia, is problematic also for more established methods, e.g., the BI adjective danger levels averaged 2.9 in these instances, as opposed to 3.2 in the rest.

Only three out of fifty instances were false negatives in Peru, most likely due to cloud cover not completely filtered in the eight-day MODIS products that we used to proxy canopy leaf moisture. The NBR and the ratio of NBR to NDMI represent the first and the third most influential predictors in the model, meaning that additional smoothing of the original 8-day MODIS products is in order. This input quality issue partly explains why the independent validation was almost perfect in drier, higher altitudes and failed more often in moist and cloudy lowlands.

4. Discussion

Most of the signal extracted from our data relates to fuel moisture, robust to fuel type misclassification, in line with the established approach to calculate static maps of fire hazard where a generic model G is used[1,3,47]. Our results relate fuel dryness with influence from wind to fire intensity outcomes, echoing recent results of systematic reviews and analysis of U.S. firefighters' entrapments and fatalities, where the ERC and BI historically highest local values effectively predict fire danger[8,9,15,16].

Our model is akin to statistical downscaling of wildfire hazard. This procedure resulted in a predictive model that improved OOS recall to 77% from the established baseline OOS recall of 61% that used only fire-danger indices as inputs. We stabilized model performance by designing our data generation process around the need to ensure high statistical power, thus with an emphasis on securing a low false negative rate out-of-sample. The independent validation conducted on official fire-related fatalities and injuries generated 30% instead of 34% false negative rate, obtained classifying the burn index (BI) with the established percentile thresholds. We identified the need to generate better optical products especially in areas with bimodal fire regimes. Those areas, corresponding to large parts of Colombia in the independent validation set, require to be explicitly included in the data generation process which in this study focused exclusively on the main fire regime of the study area comprised between July and November.

Our model can generate one reliable output: the probability that a wildfire may reach the near-maximum FRP outcome locally possible. Forecast probabilities can be aggregated at any convenient administrative scale for early warning, planning and management purposes. The model provides a hazard metric of straightforward interpretation, scaled between zero and one. Actual output values more usually vary within a smaller range oscillating around the 50% threshold. Readers can explore model outputs and MODIS fire detects ten days at a time using the second companion app to validate independent fire incidents datasets and wildfires of their interest. Both Google earth engine companion apps[60] can be accessed online[61,62].

5. Conclusions

The model we developed confirms that ERC combined with SC provide the best combination of coarse descriptors of the environment to forecast fire hazard in the study area for the sampled fire intensity according to the block maxima strategy adopted. These results might have been expected since the ERC and the SC are the components of the Burn Index, the single most important synthetic index of the US National Fire Danger Rating System (see Appendix). The sampled outputs are mostly plume dominated fires. In addition to that, we were able to correctly predict most independently collected fatal and near fatal fire incidents improving the performance of established methods by inputting finer scale descriptors of the fire environment, along with the traditional fire-danger indices, such as remotely sensed canopy characteristics, wind/topography interactions and topography.

Bimodal precipitation regimes pose a specific challenge for fire danger forecast. As readers can explore in the first companion app, most fire activities in the study area took place in clearly unimodal precipitation regimes, with one clearly defined fire-season. When canopy moisture is high year-round with two rain-seasons, and hence fire-seasons, the main driver of fire wildfire hazard as we managed to model it (fuel moisture) does not respond to the more common pattern. This subset or slice of the data is currently underrepresented in our sample. Additionally, the main predictor of fire intensity is an optical multispectral index. We used 8-day moving window MODIS products to derive the NBR, generated with cloud and shadow filters. The series is smooth but it still bumpy especially over the rainforest where the cloud cover affects detections the most. Further noise filtering could significantly improve the quality of the forecasts in these areas.

The unsatisfactory solution to the regression problem (**Figure 6**) shows one more time how important it is to always evaluate predictive models out-of-sample. Although a walk-forward cross validation is the most responsible manner to assess the reliability of a predictive model, a simple cross-validation is likely to return basically the same results, but from a larger test set, making the most of the available data. We list below a possible research agenda identified by this study to develop a fully operational fire hazard product for the area of interest:

- Explicitly include uni- and bi-modal fire regimes in the sampling procedure.
- Improve input data by filtering noise in optical products to proxy canopy characteristics.
- Use a suitable forecast product rather than reanalysis data in order to assess how a wildfire hazard operational product could perform, given the data available in real time to fire analysts.
- Develop a climatology based on a time series of fire hazard probability for adjective classification.
- Integrate the modelling of fire spreading potential to merge static and dynamic wildfire risk metrics into better operational products.

Author contributions

Conceptualization, AM and WMJ; methodology, AM and WMJ; formal analysis, AM; software, AM; validation, AM and WMJ; investigation, AM; resources, WMJ, EA and BP; data curation, AM and WMJ; writing—original draft preparation, AM and WMJ; review and editing, AM, HP, SB, CT, VAN and DRZ; visualization, AM; supervision, EA and WMJ; project administration, AM, WMJ, BP and EA; funding acquisition, EA and WMJ. All authors have read and agreed to the published version of the manuscript.

Availability of data and material

Data and results can be accessed and explored through the two web apps developed for this study. All data inputs have been derived from datasets freely available on the Google earth engine platform. Other data can be requested directly to the corresponding author.

Funding

This research was possible through the support of the United States Forest Service, Oak Ridge Institute for Science and Education, Servicio Nacional de Meteorología e Hidrología.

Acknowledgments

W. M. Jolly (USFS) supervised this work and provided scientific guidance at all time, he also generated the weather-fire indices used for this study: FWI, BI, ERC and KBDI. N. Gorelick (Google) has kindly provided useful suggestions on efficient JavaScript coding in Google earth engine. G. Lemoine (JRC) kindly provided a working code example to convert a collection of ordered multi-band images (the computed fire weather indices) into imageCollections ee objects, with a time stamp. Sampling was conducted via a JavaScript code for implementation in Google earth engine[60] and the modelling was conducted using python, relying on the scikit-learn library[56]. We would also like to thank Isidoro Solis (USFS).

Conflict of interest

The authors declare no conflict of interest.

References

1. Scott J, Thompson M, Calkin D. A wildfire risk assessment framework for land and resource management [Internet]. Gen. Tech. Rep. RMRS-GTR-315. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station; 2013. Available from: https://www.fs.usda.gov/rm/pubs/rmrs_gtr315.pdf.
2. Finney MA, McAllister SS, Grumstrup TP, *et al.* Wildland fire behaviour: Dynamics, principles and processes. CSIRO Publishing; 2021.
3. Finney M, Grenfell IC, McHugh CW, *et al.* A method for ensemble wildland fire simulation. Environmental Modeling and Assessment 2011; 16(2): 153–167. doi: 10.1007/s10666-010-9241-3.
4. Stephens SL, Burrows N, Buyantuyev A, *et al.* Temperate and boreal forest mega-fires: Characteristics and challenges. Frontiers in Ecology and the Environment 2014; 12(2): 115–122. doi: 10.1890/120332.
5. Scott AC, Holloway R, Bowman D, *et al.* Fire on Earth: An introduction. Wiley-Blackwell; 2014.
6. Preisler HK, Riley KL, Stonesifer CS, *et al.* Near-term probabilistic forecast of significant wildfire events for the Western United States. International Journal of Wildland Fire 2016; 25(11): 1169–1680. doi: 10.1071/WF16038
7. Werth PA, Potter BE, Alexander ME, *et al.* Synthesis of knowledge of extreme fire behavior: Volume 2 for fire behavior specialists, researchers, and meteorologists [Internet]. 2016. Available from: https://www.firescience.gov/projects/09-2-01-11/project/09-2-01-11_Synthesis_Extreme_FB_vol2.pdf.
8. Jolly WM, Freeborn PH. Towards improving wildland firefighter situational awareness through daily fire behaviour risk assessments in the US Northern Rockies and Northern Great Basin. International Journal of Wildland Fire 2017; 26(7): 574–586. doi: 10.1071/WF16153.
9. Jolly WM, Freeborn PH, Page WG, Butler BW. Severe fire danger index: A forecastable metric to inform firefighter and community wildfire risk management. Fire 2019; 2(3): 47. doi: 10.3390/fire2030047.
10. Podschwit H, Cullen A. Patterns and trends in simultaneous wildfire activity in the United States from 1984 to 2015. International Journal of Wildland Fire 2020; 29(12): 1057–1071. doi: 10.1071/WF19150.
11. Laurent P, Mouillot F, Vanesa M, *et al.* Varying relationships between fire radiative power and fire size at a global scale. Biogeosciences 2019; 16(2): 275–288. doi: 10.5194/bg-16-275-2019.
12. Tedim F, Leone V, Amraoui M, *et al.* Defining extreme wildfire events: Difficulties, challenges, and impacts. Fire 2018; 1(1): 9. doi: 10.3390/fire1010009.
13. Tedim F, Leone V, Mcgee T. Extreme wildfire events and disasters: Root causes and new management strategies. Elsevier; 2020.
14. Wilson CC. Fatal and near-fatal forest fires the common denominators. The International Fire Chief 1977; 43(9): 9–10.
15. Page WG, Freeborn PH, Butler BW, Jolly WM. A classification of US wildland firefighter entrapments based on coincident fuels, weather, and topography. Fire 2019; 2(4): 52. doi: 10.3390/fire2040052.
16. Page WG, Freeborn PH, Butler BW, Jolly WM. A review of US wildland firefighter entrapments: Trends, important environmental factors and research needs. International Journal of Wildland

Fire 2019; 28(8): 551–569. doi: 10.1071/WF19022.

17. Brownlee J. Imbalanced classification with Python: Better metrics, balance skewed classes, cost-sensitive learning [Internet]. Available from: https://books.google.com.co/books?id=jaXJDwAAQBAJ.
18. Ling CX, Sheng VS. Cost-sensitive learning. In: Sammut C, Webb GI (editors). Encyclopedia of machine learning. Springer; 2010. p. 231–235.
19. Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) 1958; 20(2): 215–242. doi: 10.1111/j.2517-6161.1958.tb00292.x.
20. King G, Nielsen R. Why propensity scores should not be used for matching. Political Analysis 2019; 27(4): 435–454. doi: 10.1017/pan.2019.11.
21. Iacus SM, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. Journal of the American Statistical Association 2011; 106(493): 345–361. doi: 10.1198/jasa.2011.tm09599.
22. Gertler PJ, Martinez S, Premand P, *et al.* Impact evaluation in practice. 2nd ed. Inter-American Development Bank and World Bank; 2016.
23. Good practice guidance. SDG indicator 15.3.1 [Internet]. Available from: https://www.unccd.int/sites/default/files/relevant-links/2021-03/Indicator_15.3.1_GPG_v2_29Mar_Advanced-version.pdf.
24. MapBiomas general"handbook". Algorithm theoretical basis document (ATBD) [Internet]. 2022. Available from: https://mapbiomas-br-site.s3.amazonaws.com/Metodologia/ATBD_Collection_6_v1_January_2022.pdf.
25. Algorithm theoretical base document (ATBD). RAISG-MapBiomas Amazon—Collection 4 (Spanish) [Internet]. 2022. Available from: https://s3.amazonaws.com/amazonia.mapbiomas.org/atbd/atbd%20general/ATBD_General_MapBiomas_Amazonia_4.0.pdf.
26. MapBiomas chaco general "handbook". Algorithm theoretical basis document (ATBD) [Internet]. 2020. Available from: https://mapbiomas-br-site.s3.amazonaws.com/MapBiomas%20CHACO/Colecao_2/ATBD/ATBD_-_MapBiomas_Chaco_Collection2.pdf.
27. Buchhorn M, Smets B, Bertels L, *et al.* Copernicus global land service: Land cover 100 m: Collection 3: Epoch 2015: Globe. Zenodo 2020; 1–14. doi: 10.5281/zenodo.3939038.
28. Sulla-Menashe D, Friedl MA. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product [Internet]. 2018. Available from: https://lpdaac.usgs.gov/documents/101/MCD12_User_Guide_V6.pdf.
29. Giglio L, Descloitres J, Justice CO, Kaufman YJ. An enhanced contextual fire detection algorithm for MODIS. Remote Sensing of Environment 2003; 87(2–3): 273–282. doi: 10.1016/S0034-4257(03)00184-6.
30. Lizundia-Loiola J, Otón G, Ramo R, Chuvieco E. A spatio-temporal active-fire clustering approach for global burned area mapping at 250 m from MODIS data. Remote Sensing of Environment 2020; 236: 111493. doi: 10.1016/j.rse.2019.111493.
31. Giglio L, Boschetti L, Roy D, *et al.* Collection 6 MODIS burned area product user's guide [Internet]. 2020. Available from: https://lpdaac.usgs.gov/documents/875/MCD64_User_Guide_V6.pdf.
32. Giglio L, Justice C, Boschetti L, Roy D. MCD64A1 MODIS/Terra+Aqua Burned Area Monthly L3 Global 500 m SIN Grid [Internet]. USGS. Available from: https://lpdaac.usgs.gov/products/mcd64a1v006/.
33. Giglio L, Schroeder W, Hall J, Justice C. MODIS Collection 6 and Collection 6.1 active fire product user's guide [Internet]. 2021. Available from: https://lpdaac.usgs.gov/documents/1005/MOD14_User_Guide_V61.pdf.
34. Schmit TJ, Griffith P, Gunshor MM, *et al.* A closer look at the ABI on the goes-r series. Bulletin of the American Meteorological Society 2017; 98(4): 681–698. doi: 10.1175/BAMS-D-15-00230.1.
35. Schroeder W, Csiszar I, Giglio L, *et al.* Early characterization of the active fire detection products derived from the next generation NPOESS/VIIRS and GOES-R/ABI instruments. In: Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS); 2010 Jul 25–30; Honolulu, USA. p. 2683–2686.
36. Picotte JJ, Dockter D, Long J, *et al.* LANDFIRE remap prototype mapping effort: Developing a new framework for mapping vegetation classification, change, and structure. Fire 2019; 2(2): 35. doi: 10.3390/fire2020035.
37. Reeves MC, Ryan KC, Rollins MG, Thomas TG. Spatial fuel data products of the LANDFIRE project. International Journal of Wildland Fire 2009; 18(3): 250–267. doi: 10.1071/WF08086.
38. Hansen MC, Potapov PV, Moore R, *et al.* High-resolution global maps of 21st-century forest cover change. Science 2013; 342(6160): 850–854. doi: 10.1126/science.1244693.
39. Alexander ME, Cruz MG. Fireline intensity. In: Manzello S (editor). Encyclopedia of wildfires and wildland-urban interface (WUI) fires. Springer; 2018. p. 1210.
40. Cheng L, Yajun L, Chang Z, *et al.* The method of evaluating sub-pixel size and temperature of fire spot in AVHRR data. Journal of Applied Meteorological Science 2004; 15(3): 273–280.
41. Peterson D, Wang J, Ichoku C, Hyer EJ. Sub-pixel fractional area of wildfires from MODIS observations: Retrieval, validation, and potential applications. In: Proceedings of the 34th Interna-

tional Symposium on Remote Sensing of Environment—The GEOSS Era: Towards Operational Environmental Monitoring; 2011 Apr 10–15; Sydney, Australia.

42. Gumbel EJ. Statistics of extremes [Internet]. Available from: https://mathscinet.ams.org/mathscinet/relay-station?mr=0096342.
43. Muñoz-Sabater J, Dutra E, Agustí-Panareda A, *et al*. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data 2021; 13(9): 4349–4383. doi: 10.5194/essd-13-4349-2021.
44. Lawson BD Armitage OB. Weather guide for the Canadian forest fire danger rating system [Internet]. Available from: https://cfs.nrcan.gc.ca/pubwarehouse/pdfs/29152.pdf.
45. Jolly WM, Cochrane MA, Freeborn PH, *et al*. Climate-induced variations in global wildfire danger from 1979 to 2013. Nature Communications 2015; 6: 7357. doi: 10.1038/ncomms8537.
46. Keetch JJ, Byram GM. A drought index for forest fire control [Internet]. U.S.D.A. Forest Service Research Paper SE - 38; 1968. Available from: https://www.srs.fs.usda.gov/pubs/rp/rp_se038.pdf.
47. Heinsch FA, Andrews PL, Tirmenstein D. How to generate and interpret fire characteristics charts for the U.S. fire danger rating system [Internet]. Gen. Tech. Rep. RMRSGTR-363. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station; 2017. Available from: https://www.fs.usda.gov/rm/pubs_series/rmrs/gtr/rmrs_gtr363.pdf.
48. Jolly WM, Butler BW, Forthofer J. Assessing topography and wind alignment for firefighter safety [Internet]. Available from: file:///C:/Users/Administrator/Downloads/156361.pdf.
49. Huang S, Tang L, Hupy JP, *et al*. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. Journal of Forestry Research 2020; 32(5): 1–6. doi: 10.1007/s11676-020-01155-1.
50. Keeley JE. Fire intensity, fire severity and burn severity: A brief review and suggested usage. International Journal of Wildland Fire 2009; 18(1): 116–126. doi: 10.1071/WF07049.
51. Gao BC. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 1996; 58(3): 257–266. doi: 10.1016/S0034-4257(96)00067-3.
52. Armenteras D, Dávalos LM, Barreto JS, *et al*. Fire-induced loss of the world's most biodiverse forests in Latin America. Science Advances 2021; 7(33): 2–10. doi: 10.1126/sciadv.abd3357.
53. Barreto JS, Armenteras D. Open data and machine learning to model the occurrence of fire in the ecoregion of "Llanos Colombo–Venezolanos". Remote Sensing 2020; 12(23): 3291. doi: 10.3390/rs12233921.
54. Jain P, Coogan SCP, Subramanian S, *et al*. A review of machine learning applications in wildfire science and management. Environmental Reviews 2020; 28(3): 73. doi: 10.1139/er-2020-0019.
55. Suradhaniwar S, Kar S, Durbha S, Jagarlapudi A. Time series forecasting of univariate agrometeorological data: A comparative performance evaluation via one-step and multi-step ahead forecasting strategies. Sensors 2021; 21(7): 2430. doi: 10.3390/s21072430.
56. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 2011; 12(85): 2825–2830. doi: 10.1289/EHP4713.
57. Boehmke B, Greenwell BM. Hands-on machine learning with R. 1st ed. Chapman and Hall/CRC; 2019.
58. Ledolter J, Kardon RH. Focus on data: Statistical design of experiments and sample size selection using power analysis. Investigative Ophthalmology and Visual Science 2020; 61(8): 11. doi: 10.1167/IOVS.61.8.11.
59. Alzate DF, Carrillo GAA, Barbosa EOR, *et al*. REGNIE interpolation for rain and temperature in the Andean, Caribbean, and Pacific regions of Colombia. Colombia Forestal 2018; 21(1): 102–118. doi: 10.14483/2256201X.11601.
60. Gorelick N, Hancher M, Dixon M, *et al*. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 2016; 202; 18–27. doi: 10.1016/j.rse.2017.06.031.
61. Earth Engine Apps. Available from: https://anmarkos.users.earthengine.app/view/forecasting-wildfire-hazard-across-nw-south-america-app1.
62. Earth Engine Apps. Available from: https://anmarkos.users.earthengine.app/view/forecasting-wildfire-hazard-across-nw-south-america-app2.

Appendix

Fire characteristics chart for the U.S. fire danger rating system

The energy release component is considered a composite fuel moisture index as it reflects the contribution of all live and dead fuels to potential fire intensity. Each daily calculation considers the past seven days in calculating the new number. Daily variations of the ERC are small as wind is not part of the calculation. The ERC is used as a tool for daily staffing and firefighters' situational awareness. NFDRS 1978 fuel model G is widely used to display ERC as it contains all the dead size class fuels and both the herbaceous and woody live fuels[47]. The spread component uses the original weighting factors, emphasizing the fine fuels that carry fire spread. Unlike ERC-G, heavy fuels are not included in the SC calculation, SC can vary greatly from one day to the next[47] (**Figure A1**).

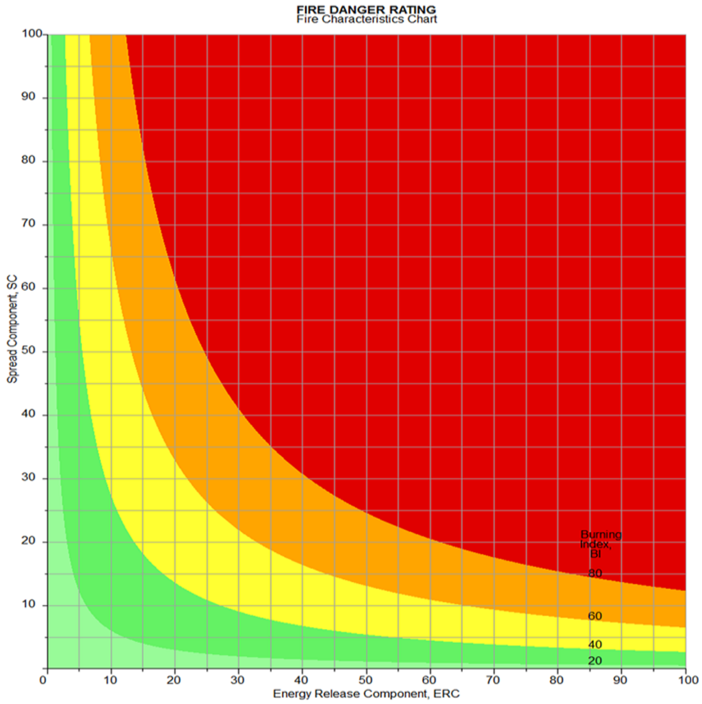

Figure A1. Fire characteristics chart for the U.S. fire danger rating system.