CSIRO PUBLISHING

International Journal of Wildland Fire **2020**, *29*, 11–27
https://doi.org/10.1071/WF19058

Mesoscale spatiotemporal predictive models of daily human- and lightning-caused wildland fire occurrence in British Columbia

Khurram Nadeem[A,*], *S. W. Taylor*[B,E,*], *Douglas G. Woolford*[C] *and C. B. Dean*[D]

[A]University of Guelph, 50 Stone Road E, Guelph, ON N1G 2W1, Canada.
[B]Pacific Forestry Centre, Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada.
[C]University of Western Ontario,1151 Richmond Street, London, ON N6A 3K7, Canada.
[D]University of Waterloo, 200 University Avenue W, Waterloo, ON N2L 3G1, Canada.
[E]Corresponding author. Email: steve.taylor@canada.ca

Abstract. We developed three models of daily human- and lightning-caused fire occurrence to support fire management preparedness and detection planning in the province of British Columbia, Canada, using a lasso-logistic framework. Novel aspects of our work involve (1) using an ensemble of models that were created using 500 datasets balanced (through response-selective sampling) to have equal numbers of fire and non-fire observations; (2) the use of a new ranking algorithm to address the difficulty in interpreting variable importance in models with a large number of covariates. We also introduce the use of cause-specific average spatial daily fire occurrence, termed baseline risk, as a covariate for missing or poorly estimated factors that influence human and lightning fire occurrence. All three models have strong predictive ability, with areas under the Receiver Operator Characteristic curve exceeding 0.9.

Additional keywords: fire danger, fire occurrence modelling.

Received 9 April 2019, accepted 16 October 2019, published online 24 December 2019

Introduction

The high interannual variation in lightning-caused fires in western North America due to intense lightning storms in some years is well known (Show and Kotok 1923; Melrose and Holmgren 1932). In the Province of British Columbia (BC), Canada, considerable variation in both the number of daily lightning-caused (mean 30, range 0–400) and human-caused (mean 10, range 0–30) wildfires and their spatial distribution is due to complex interactions between the occurrence of cloud-to-ground lightning strikes and a variety of human activities with synoptic-scale influences of atmospheric circulation, mesoscale influences of complex topography, and microscale influences of diverse vegetation types on fuel flammability. Strong spatial structure in the location of human-caused fires in BC is associated with settlement and development patterns (e.g. roads, railways, recreation and industrial activity), which are constrained by rugged topography (Camp and Krawchuk 2017). Seasonal trends in human-caused fires are also influenced by seasonal variation in vegetation phenology and human activity, as well as fuel flammability.

In order to effectively plan the types, amount, positioning and readiness of resources that may be needed to respond to fires expected in the upcoming days, fire managers need a spatially explicit estimate of the daily fire load. Substantial variation in the number of new fires that occur each day or over a few days, and especially surges in lightning-caused fire starts, has long been recognised as a major fire management planning challenge (Hornby 1936) and has motivated work on fire danger rating. Early work on fire prediction examined the relationship between fire occurrence and individual meteorological variables such as relative humidity on seasonal (Saari 1923) and daily bases (Noble 1926). Subsequently, fire danger indices that were developed to incorporate the cumulative effects of multiple meteorological variables (e.g. temperature, precipitation, relative humidity) into indicators of fuel flammability and fire behaviour (Taylor and Alexander 2006; Hardy and Hardy 2007) enabled analysis of the empirical frequency of fire occurrence by fire danger class (Beall 1934), or the expected occurrence of one or more fires by danger class by administrative region (Crosby 1954).

Fire occurrence patterns are inherently random. Consequently, stochastic and statistical models that incorporate random variables into their structure are natural frameworks for modelling wildland fire occurrence. A very early stochastic

[*]KN and SWT contributed equally to the paper.

framework for predicting fire occurrences was the negative binomial model for new fire counts as a function of a fire danger index (Bruce 1963). Later, Cunningham and Martell (1973) used a Poisson model to relate counts of fires to the Fine Fuel Moisture Code (FFMC), a component of the Canadian Forest Fire Weather Index (FWI) System (see Van Wagner 1987; or Wotton 2009). In the following 45 years, statistical modelling evolved along with increasing data availability. Remote automated fire weather stations developed in the 1980s have generated 30 years of weather observations from many more locations than are represented in national meteorological networks. Advances in numerical weather modelling have also facilitated development of gridded reanalysis datasets. Lightning location detectors developed in the 1980s (Noggle *et al.* 1976 have produced decades of strike location data (Gilbert and Zala 1987), while an increasing volume of data on vegetation properties including seasonal greenness has been accrued from satellite-borne sensors.

With higher resolution of fire-weather data has come more sophisticated modelling methods. Taylor *et al.* (2013) described a well-established methodology as introduced in the seminal work of Brillinger *et al.* (2003), namely, a discretised approach to modelling fire occurrence with multiple covariates where the underlying fire occurrence process is assumed to be a spatiotemporal point process with an inhomogeneous conditional intensity function that depends on a variety of predictors; these may include local weather and fuel moisture conditions as well as other key variables such as local land-cover and a land-use characteristics. Typically, logistic generalised additive models are used in this framework where non-linear relationships between the log-odds of fire occurrence risk and predictors are modelled using spline-based smoothers. Representative examples of this modelling technique include Brillinger *et al.* (2003), Vilar *et al.* (2010) and Woolford *et al.* (2010). For further technical details and references, see the reviews in Taylor *et al.* (2013) and Xi *et al.* (2019).

Despite these advances, several challenges remain. These include: (1) the collection, fusion and alignment of various sources of meteorological, geographic and demographic data of different source resolutions that are required to assemble daily values of candidate covariates over several fire seasons on a fine spatial grid (often in the order of 1–400 km^2), generating large datasets; (2) responses may be non-linear with complex interactions between variables; (3) discretisation of the spatial domain into finer scales leads to large class imbalance between fire and non-fire events whereby fire occurrence becomes an uncommon if not rare-event problem; (4) it is difficult to measure the intensity of human activity that could result in a fire at fine spatiotemporal scales; and (5) although many lightning-caused fire occurrence models and over 200 person-caused fire models have been developed in the past 45 years (Costafreda-Aumedes *et al.* 2017), relatively few models are implemented operationally by fire management organisations.

In Canada, surface weather observations from ~2000 weather stations, as well as lightning strike locations, are incorporated in the Canadian Wildland Information System (CWFIS, Lee *et al.* 2002) at a national scale and in provincial and territorial fire management agency information systems in near-real time. Surface and atmospheric forecast data up to 14 days are also incorporated in CWFIS from the North American Ensemble Forecast System (NAEFS, Toth *et al.* 2005). These systems provide a good framework for implementing fire occurrence models.

Objectives

The objectives of our research are to: (i) describe the methods and procedures involved in assembling a large database for spatiotemporal modelling of fire occurrence; (ii) develop a modelling framework to determine key driving factors of human- and lightning-caused fires; and (iii) develop three models of daily lightning- and human-caused fires that can be readily implemented within fire management information systems for different management applications:

- An Observed Lightning-Caused Fire (OLCF) model that can be used to nowcast[1] fires that have occurred (but may or may not have yet been reported) informed in part by recent lightning strike and weather observations;
- A Predicted Lightning-Caused Fire (PLCF) model to forecast lightning fires, informed in part by weather and atmospheric stability measures that can be calculated from medium-term numerical weather model forecasts;
- A Human-Caused Fire (HCF) model to that can be used to nowcast and forecast human-caused fires informed by observed or forecast surface weather conditions.

Earlier work on daily fire prediction models for BC by Magnussen and Taylor (2012) utilised data from 1970 to 2000. Here, we introduce several novel analytical procedures and statistical modelling methods. In the *Variable selection and data compilation* subsection, we introduce human- and lightning-caused fire baseline risk covariates based on ranking of the average number of historical fires per grid cell in the available incident database since 1981. The baseline risk was employed to account for spatial variation in fire occurrence that is not well explained by specific geographic variables, e.g. forest cover, road density. It is analogous to the incidence of disease in a population in epidemiology (e.g. Wang *et al.* 2009) that anchors the event risk. This quantity was termed persistence probability by Preisler and Westerling (2007), who used it to illustrate anomalies. We also use a measure of vegetation greenness (Normalized Difference Vegetation Index, NDVI) derived from remote sensing to explain seasonal variation in fire occurrence not fully accounted for by daily weather and fire danger measures, and use several atmospheric stability indices derived from a reanalysis dataset as indicators of lightning strike potential. In the *Response-based sampling* subsection, we address the class imbalance problem (>99% of voxels have no fire) by using sampling to create balanced datasets with the same number of fire and no-fire observations, and in the

[1] Two separate lightning-caused fire models are needed for nowcast and forecast applications We define nowcast as a prediction of the probability that a fire was ignited at a location in the recent past, which may or may not have yet been reported, using a model that includes observed weather and lightning strikes. Such models can be used to guide detection efforts following a lightning storm.

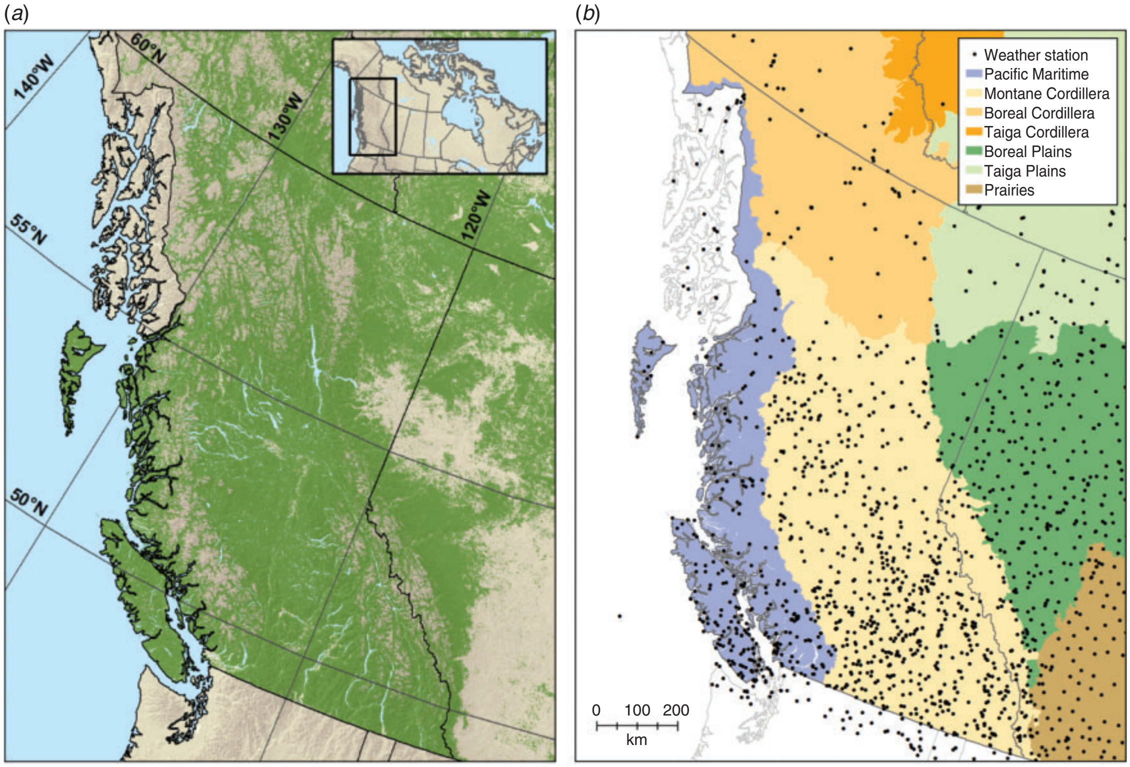

Fig. 1. (*a*) Coniferous forest cover in the study area (green shading) and location within Canada (inset). (*b*) Ecoregions in the study area and locations of stations used in compiling the historic weather dataset.

Variable ranking subsection, we introduce ensemble methods to rank variable importance. A list of abbreviations used in the paper is in the Supplementary material available on the journal's website.

Materials and methods

In this section, we describe six phases in our model development: variable selection and data compilation, model selection, response-based sampling, model fitting, model evaluation and variable ranking. Prior to doing so, we first briefly review the influences of weather, topography, vegetation and ignition sources on fuel flammability and fire occurrence in BC, which informed variable selection.

Fire environment of the study area

The model domain is the Canadian province of British Columbia, which has a land area of 945 000 km^2 falling between 48° and 60°N latitude. Approximately 70% of the land area is dominated by coniferous forest or grassland and is potentially flammable (Fig. 1*a*).

The moisture content and flammability of forest fuels in BC is strongly affected by the position of the mid-latitude storm track that delivers moisture-laden Pacific air to western North America, which is closely connected to the strength and position of alternating low and high stratospheric pressure features (Moore *et al.* 2010). The Aleutian Low (AL) centred in the Gulf of Alaska is the dominant synoptic feature in winter (Stahl *et al.* 2006). The North Pacific High (NPH) is the dominant synoptic feature in during the April–October fire season, moving north and increasing in intensity and persistence in summer. The NPH blocks the flow of moisture-laden air, resulting in warm temperatures and rain-free periods of several days to many weeks in duration (especially in July and August), and a trend to increasing drying of deep organic layers over the fire season. The blocking effect is strongest in southern BC, but can extend over the province. Although the AL moves north and weakens in summer, weak low-pressure systems connected to the AL intrude periodically and may bring frontal precipitation over widespread areas during the fire season. The North American Cordillera that runs through the province from north to south further influences mesoscale variation in weather and climate. The various mountain ranges comprising the Cordillera interrupt the west-to-east zonal flow and restrict the westward flow of continental and Arctic air masses from central and northern Canada, resulting in rain shadow effects on the east side of the mountain ranges, and a strong west–east gradient in increasing temperature, and decreasing precipitation (Moore *et al.* 2010). However, greater atmospheric instability east of the Coast Mountains results in periodic local convective storms with precipitation in summer months (Jackson 1968).

A variety of ecosystems ranging from temperate rainforests to semideserts and grasslands, boreal forests and alpine tundra have developed in BC, varying with proximity to the Pacific Ocean, topography and latitude (Meyn *et al.* 2010). These ecosystems have different structural features, such as canopy

Table 1. Covariates used in the daily fire occurrence prediction models

WUI, wildland–urban interface; WII, wildland–industrial interface; NDVI, Normalized Differential Vegetation Index

Variable	Definition	Variable	Definition
Baseline risk		**Surface fire weather**	
LIGHTNING RISK RANK	Ranked lightning fire occurrence rate	TEMPERATURE	Temperature at noon
LOGIT HUMAN RISK	Logistic transform of human fire risk occurrence rate	RELATIVE HUMIDITY	Relative humidity at noon
LOGIT HUMAN RISK^2	Square of LOGIT HUMAN RISK	WIND SPEED	10-min avg wind speed at noon
Geographic		PRECIPITATION	24-h precipitation at noon
LATITUDE	Latitude of cell midpoint	FFMC	Fine Fuel Moisture Code
LONGITUDE	Longitude of cell midpoint	DMC	Duff Moisture Code
ELEVATION	Mean cell elevation	SDMC	Sheltered Duff Moisture Code
ROUGHNESS	Standard deviation of elevation	DC	Drought Code
ELEVATION^2	Square of ELEVATION	ISI	Initial Spread Index
BOREAL CORDILLERA	Ecoregion of cell (0,1)	BUI	Build-up Index
BOREAL PLAIN	Ecoregion of cell (0,1)	FWI	Fire Weather Index
PACFIC MARITIME	Ecoregion of cell (0,1)	DSR	Daily Severity Rating
TAIGA PLAIN	Ecoregion of cell (0,1)	PSUF	Probability of sustained flaming ignition
MONTANE CORDILLERA	Ecoregion of cell (0,1)	TEMP^2	Square of temperature
Time periods		DMC^2	Square of DMC
CHANGE POINT	1 if year >1992, else 0	DC^2	Square of DC
WEEKDAY$_X$	1 if weekday = X, else 0	ISI^2	Square of ISI
Vegetation		FWI^2	Square of FWI
VEGETATED	Vegetated proportion	PRECIPITATION LAG1	Precipitation at day $_{t-1}$
TREED	Treed proportion	PRECIPITATION LAG2	Precipitation at day $_{t-2}$
CONIFER COVER	Proportion of conifer species	PRECIPITATION LAG3	Precipitation at day $_{t-3}$
DECIDUOUS COVER	Proportion of deciduous species	ACCUM PRECIPITATION	Precipitation in days $_{(t\ldots t-3)}$
%CONIFER	Percentage of treed area conifer	FFMC × TEMPERATURE	FFMC × temperature
%DECIDUOUS	Percentage of treed area deciduous	DMC × TEMPERATURE	DMC × temperature
%MIXEDWOOD	Percentage of treed area mixed wood	DC × TEMPERATURE	DC × temperature
AVERAGE NDVI	Mean NDVI value per day-cell	ISI × TEMPERATURE	ISI × temperature
%CONIFER^2	Square of % CONIFER	BUI × TEMPERATURE	BUI ×temperature
%DECIDUOUS^2	Square of % DECIDUOUS	FWI × TEMPERATURE	FWI × temperature
AVERAGE NDVI^2	Square of AVERAGE NDVI	DC..ACCUM PRECIP	DC/(1 + ACCUM PRECIPITATION)
Ecumene		DMC..ACCUM PRECIP	DMC/(1 + ACCUM PRECIPITATION
ROAD LENGTH	Sum of road segment lengths	FFMC..ACCUM PRECIP	FFMC/(1 + ACCUM PRECIPITATION)
POPULATION	Population density	ISI..ACCUM PRECIP	ISI/(1 + ACCUM PRECIPATION)
WUI AREA	WUI area within each cell	BUI..ACCUM PRECIP	BUI/(1 + ACCUM PRECIPITATION)
WII AREA	WII area within each cell	FWI..ACCUM PRECIP	FWI/(1 + ACCUM PRECIPITATION)
WUI DISTANCE	Distance to nearest WUI polygon	**Atmospheric stability**	
WII DISTANCE	Distance to nearest WII polygon	500 MB ANOMALY	500 mb (hPa) geopotential height anomaly
ROAD LENGTH$^{0.5}$	Square root of ROAD LENGTH	500 MB TENDENCY	500 mb geopotential ht day t – (t – 1)
POPULATION$^{0.5}$	Square root of POPULATION	K INDEX	K Index
Lightning		TOTALS INDEX	Totals Index
LIGHTNING STRIKES	Cloud-to-ground lightning strikes counted in the previous 24-h	SHOWALTER INDEX	Showalter Index
LIGHTNING LAG1	Lightning strikes in day $_{t-1}$	C-HAINES INDEX	Continuous Haines Index
LIGHTNING LAG2	Lightning strikes in day $_{t-2}$		
ACCUM LIGHTNING	Lightning strikes in days $_{(t\ldots t-3)}$		
LIGHTNING INDICATOR	1 if ACCUM LIGHTNING >0, else 0		

closure, that influence fuel wetting and drying rates, and different surface fuel properties such as organic layer depth, that affect moisture-holding capacity. Five broad ecozones are recognised: Pacific Maritime, Montane Cordillera, Boreal Cordillera, Boreal Plains and Taiga Plains (Fig. 1*b*).

Lightning may be produced both by the passage of upper low troughs and convective storms (Alexander 1927). Individual events comprising thousands of strikes have a strong spatial structure associated with storm tracks. Lightning strike density increases approximately along the west–east gradient in continentality (Burrows *et al.* 2002).

Human activity in BC's forests (represented by population and road density) generally decreases along a south–north gradient, and is concentrated in valley bottoms.

Variable selection and data compilation

Our models represent the average concurrent relationship between a set of explanatory variables (represented by the data

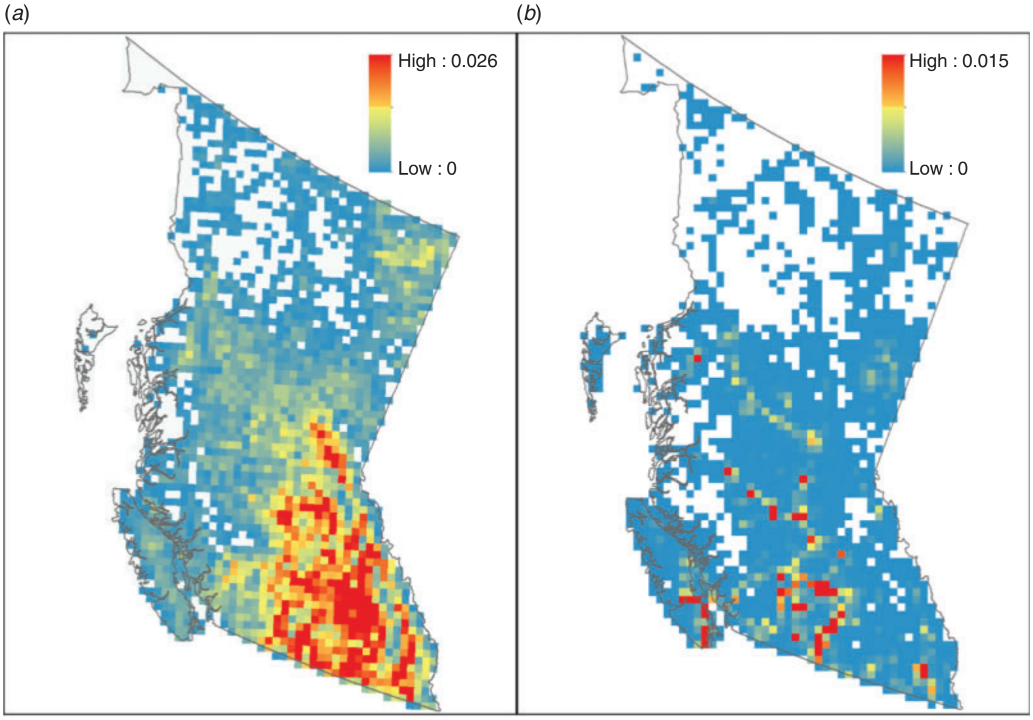

Fig. 2. (*a*) Baseline lightning-caused fire ignition rate (fires day^{-1} 400 km^{-2}); (*b*) baseline person-caused fire ignition rate (fires day^{-1} 400 km^{-2}). No fires were detected in the white cells during the period 1985–2014. The aggregate background rate for all British Columbia was ~1700 fires per year or 4.1 and 3.8 person- and lightning-caused fires per day respectively during 10 March–15 October.

matrix X) and Bernoulli fire occurrence (represented by the response vector Y). The sampling unit is a space-time voxel that represents a 24-h time period (starting at midnight) for a 20 × 20-km cell in the National Forest Inventory (NFI) grid (Gillis *et al.* 2005). The grid comprises 2 541 400-km^2 spatial units in BC. The spatial resolution was chosen as a compromise between a desire for high spatial resolution and issues with weather data accuracy and increasing rarity of fire occurrences at fine scales. The temporal domain of the models was restricted to the fire season, a 7-month period from 16 March to 14 October, which captures more than 99% of fires that occurred in BC.

Based on observation and understanding of weather and fire dynamics in BC, experience from related modelling work and preliminary analyses, we considered 72 and 83 of the variables listed in Table 1 as candidate explanatory variables for the lightning- and human-caused models respectively. They include spatially varying but temporally static baseline, geographic, vegetation and ecumene variables, and spatiotemporally varying lightning strike and meteorological variables, which along with fire occurrence records, are a mixture of point, interval and continuous data in their raw form. All explanatory variables were binned by day and cell or interpolated to the centroid of each 20 × 20-km grid cell, as described in the following sections. All daily values are assumed to represent a 24-h period (0000–2400 hours). As many as 18 million voxels were created for each variable (2541 cells × 34 years × 214 days) in the data compilation phase. The compilation and management of these the primary and gridded covariate datasets, although time-consuming, is a necessary task. Note that M random subsets of voxels from this large and complex spatiotemporal data structure were used for model fitting, as described in the *Response-based sampling* section to follow.

Fire response and baseline occurrence risk

Records of characteristics (start date, cause, latitude, longitude) of ~70 000 individual fires were obtained from the BC Wildfire Service for the 1980–2014 period, and then mapped to our space–time voxels. At the spatial resolution of our analysis, most of counts of fires are reduced to either 0s or 1s, which represent non-fire and fire events respectively. This discretised approach permits the use of logistic modelling methods to estimate fire occurrence probability.

The baseline risk of lightning-caused and human-caused fire occurrence was estimated as the observed average daily count of fire occurrences in a cell determined using the locations of individual georeferenced fires. These rates are point estimates of long-term expectations of the response variable over the spatial grid (Fig. 2) and enter into the models as transformed covariates (lightning risk rank and logit human risk in Table 1*a*).

Geographic

Ten variables that may contribute to fixed topographical or ecoloclimate effects were evaluated. The ecozones of Canada (Wiken 1986) were used to regionalise the models; cell values were obtained by rasterising a polygon-based map (Fig. 1*b*). Cells in the same ecozone are assumed to have similar ecological characteristics. Digital elevation data over Canada were obtained at 7.3 arc s (225-m) grid resolution from the GMTED2010 – Global MultiResolution Terrain Elevation Dataset distributed by the US Geological Survey (http://earthexplorer.usgs.gov/, accessed 12 November 2019) and upscaled to a mean and standard deviation for the 20-km cells, where standard deviation of the elevation is an index of topographic roughness.

Vegetation

Fifteen measures of vegetation cover were obtained from a national forest inventory dataset developed by Beaudoin *et al.* (2014) by imputing various vegetation properties obtained from aerial photo and ground plots to a 250-m Moderate Resolution Imaging Spectroradiometer (MODIS)-based grid using *k*-nearest neighbour procedures. We determined the mean and standard deviation of the proportion of vegetated and non-vegetated area; the proportion of treed and non-treed area; and the proportion of coniferous and deciduous tree species by upscaling the 250-m grid to our 20-km grid. In addition, we estimated the proportion of the treed area in each 20-km cell that was conifer, deciduous or mixed-wood classes (>75, 26–75, and <25% needled leaf proportion respectively) as the empirical proportion of the number of 250-m cells in each needle-leafed class and the number of treed 250-m cells in a 20-km cell.

The timing and intensity of spring leaf flush and fall (autumn) senescence vary geographically and interannually and influence surface fuel flammability in temperate and boreal forests but are not represented in the current Canadian Forest Fire Danger Rating System. After the snow melts in spring, there may be a period of several weeks before understorey vegetation flushes when there is an accumulation of flammable dead vegetation on the ground surface, and in deciduous and mixed-wood forests, a period when there is greater insolation and wind penetration within forest stands before trees leaf out. Leaf flush typically occurs over a period of weeks, with leaf cover reaching a plateau in late spring to early summer, remaining fairly constant for the growing season before senescence begins again in fall. The NDVI is a transformation of satellite-based spectral reflectance measurements acquired in the visible (red) and near-infrared regions, which vary with vegetation cover. We obtained historical daily NDVI data acquired by the AVHRR satellite from the NOAA National Climatic Data Centre (Vermote *et al.* 2014) for the BC domain from 1981 to 2015. The native data at 250-m resolution were upscaled to 20 km. We then estimated mean NDVI values for each cell and day-of-year combination by fitting spline curves to the time series of averaged daily NDVI values. The resulting covariate, mean NDVI, serves as an average phenological index for a grid cell.

Ecumene

The potential for human-caused fire ignitions (that may result in a fire report) within the ecumene or inhabited portion of BC is difficult to assess directly: eight proxy measures were evaluated in addition to baseline risk. The total length of roads (km) in each cell, obtained from a national road database (Statistics Canada 2015), was used as a proxy for accessibility. Population counts were also obtained for census dissemination block polygons (the smallest spatial unit for which population data are available) in the 2011 census (Statistics Canada 2012) and an areal interpolation algorithm (districting) was used to estimate the population in a grid cell (de Smith *et al.* 2015). Furthermore, the proportion of a cell composed of, or the distance from the midpoint to a wildland–urban interface (WUI) or wildland–industrial interface (WII) features was also included (Johnston and Flannigan 2018). Days of the week were also included as a binary categorical variable (e.g. MONDAY = 0, 1).

Lightning strikes

We obtained data on individual lightning strike locations and strike times from the Canadian Lightning Detection Network (CLDN) for 1998–2015 (Dockendorff and Spring 2005) and binned the number of strikes per day, and lagged counts for 1 and 2 days and accumulated counts for 3 days for each cell.

Surface weather and fire danger indices

Thirty-two measures of surface weather, fuel moisture, and fire danger and transforms were evaluated. Historical daily observations of temperature, relative humidity, wind speed and 24-h precipitation at 1200 hours were obtained for Meteorological Service of Canada (MSC) and provincial and territorial fire management agency stations (including BC) in Canada as well as for National Weather Service (NWS) and US fire agency remote automated weather stations (RAWS) stations in the adjacent US states within 60 km of the border for the 1980–2014 period (Fig. 1*b*). The MSC and NWS stations have continuous daily observations, whereas many of the provincial and territorial and US RAWS stations are only operational during the fire season. Daily observations of these variables were also obtained from the North American Regional Reanalysis (NARR) dataset (Mesinger *et al.* 2006) at 0.3° resolution (~32 km grid) and interpolated to the 20-km grid using thin plate spline (TPS) regression technique as implemented in the *R* package *fields* (Nychka *et al.* 2017). Station-based weather variables were also interpolated to the 20-km grid using TPS with corresponding NARR-based interpolant serving as a covariate, and elevation as an additional covariate for temperature. The NARR data provide information on spatial variation in these variables at a daily scale related to synoptic-scale weather patterns, modifying the influence of distance and elevation. Cross-validation analysis revealed that including NARR data improved interpolation, especially in areas of low station density, and reduced border effects.

Although there is an operationally defined fire season across the province of BC, a given location (e.g. a grid cell) is not at risk of fire occurrence until certain conditions occur. Fire season start dates are not stationary and were calculated for each cell for each year using two rules, depending on the amount of tree cover per cell. For open cells (tree cover <75%), the fire season start day was the last of day of the series of 5 consecutive days >9°C (Simard and Valenzuela 1972) after 10 March. For densely forested cells (tree cover >75%) where snow persists longer, the fire season start day was the last of day of the series of 5 consecutive days >10.5°C after March 10th. Fire season end dates were calculated as the date when noon temperature <5°C for 3 consecutive days (Wotton and Flannigan 1993). Overwinter precipitation was calculated from the fire season end date to start date of successive years. The six values Fine Fuel Moisture Code, Duff Moisture Code, Drought Code, Initial Spread Index, Buildup Index and Fire Weather Index (FFMC, DMC, DC, ISI, BUI, FWI, respectively) of the FWI System and the sheltered duff moisture code, SDMC (Wotton *et al.* 2005), were calculated for the fire season days via the *R* package *cffdrs* (Wang *et al.* 2017) using standard values to initialise the calculations at the beginning of the fire season, with modification of the DC value according to the over-winter precipitation (Lawson and Armitage 2008. FWI System values outside the fire season dates are recorded as NA (not applicable).

Atmospheric stability

Six atmospheric stability indices were calculated from formulae constructed to reflect the potential for convection within an air mass from temperature and humidity measures at certain fixed critical levels in the atmosphere. The K, Showalter and Totals Indices, indicators of lightning storm development, were calculated following Stull (2015). The continuous Haines Index, a measure of potential for 'blow-up' fire conditions, was calculated after Mills and McCaw (2010). Daily temperature and dew point temperature estimates at 850, 750 and 500 hPa (mb) needed to calculate these four indices were obtained from the National Centre for Environmental Prediction (NCEP) reanalysis dataset (Kalnay *et al.* 1996) on a 2.5° × 2.5° grid (approx. 200 × 200 km) and re-interpolated to our 20-km grid. NCEP 500-mb geopotential height data were also used to calculate the daily 500-mb geopotential height anomaly as the difference from the average height of the 500-mb pressure level on each day in each cell, and the 500-mb height tendency as the difference between sequential days. Negative 500-mb anomalies and tendencies associated with frontal passage have been associated with lightning ignitions and extreme fire behaviour in Alberta (Nimchuk 1983; Janz and Nimchuk 1985).

Model selection: the lasso-logistic model

Binary response outcomes of fire occurrence on a given day–cell combination, (Y) such as zero ($Y = 0$) or at least one ($Y = 1$) can be modelled using the ordinary logistic regression (OLR) model with the joint likelihood function for n observations given as: $L = \prod_n \pi^y (1-\pi)^{1-y}$, where y is the observed fire occurrence and $(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{\beta}^t \underline{\boldsymbol{x}}$, where $\boldsymbol{\beta}$ is the vector of regression coefficients $\boldsymbol{x}$ including the intercept, $\underline{\boldsymbol{x}} = (1, \boldsymbol{x})^t$; and t denotes vector transpose. The resulting log-likelihood function can be expressed as follows:

$$l(\boldsymbol{\beta}) = -\sum\nolimits_{i=1}^{n} [(1-y_i)\boldsymbol{\beta}^t \underline{\boldsymbol{x}}_i + \ln(1 + \exp(-\boldsymbol{\beta}^t \underline{\boldsymbol{x}}_i))]. \qquad (1)$$

Here, we model fire occurrences using the lasso-logistic regression model (Tibshirani 1996), which is a regularised version of OLR with an L_1-penalty imposed on the coefficients vector $\overline{\boldsymbol{\beta}} = (\beta_1, \beta_1, \ldots, \beta_p)^t$, i.e. we now maximise Eqn 1 subject to the constraint C that $\sum_{k=1}^{P} |\beta_k| \leq C$; $C > 0$ (k is an index of β_k). This leads to the following *penalised* form of Eqn 1:

$$\begin{aligned} l_1(\boldsymbol{\beta}) = &-\sum\nolimits_{i=1}^{n} [(1-y_i)\boldsymbol{\beta}^t \underline{\boldsymbol{x}}_i + \ln(1 + \exp(-\boldsymbol{\beta}^t \underline{\boldsymbol{x}}_i))] \\ &- \lambda \sum\nolimits_{k=1}^{P} |\beta_k|, \end{aligned} \qquad (2)$$

where $\lambda \geq 0$ is the Lagrangian of the optimisation problem in Eqn 2. This form of the penalised likelihood function can also be derived under a double-exponential (Laplacian) prior distribution on $\overline{\boldsymbol{\beta}}$ in a hierarchical Bayesian formulation of OLR (Lee *et al.* 2006). The tuning parameter λ can be estimated using various methods including generalised and k-fold cross-validation (Tibshirani 1996). We employ the latter approach in fitting lasso-logistic model to BC data using the *R* package *glmnet* (Friedman *et al.* 2010).

We fit three separate lasso-logistic models regression models that have different ignition indicators for different management applications (see *Objectives* subsection): (i) an OLCF model of lightning-caused fire occurrence that includes observed lightning strikes; (ii) a PLCF model of lightning-caused fire occurrence that excludes lightning strike covariates but includes atmospheric stability indices; and (iii) an HCF model including indicators of human activity. Lists of covariates associated with these models are reported in Supplementary Tables S1–S3.

Response-based sampling

Wildland fire occurrences are exceedingly rare at fine space–time resolutions. We have ~18 million 20 km × 20 km × 1 day space–time voxels spanning 34 fire seasons (1981–2014); lightning and human-caused fire occurrences ($Y = 1$) were recorded in only 0.23 and 0.18% of these voxels respectively. In order to resolve the resulting class-imbalance (see, for an overview, Haixiang *et al.* 2017) in response and the computational intractability of fitting models to prohibitively large datasets, we employed a response-based *downsampling* scheme for controls ($Y = 0$), where all case observations are retained together with a simple random sample drawn from controls. This sampling procedure is akin to epidemiological case-control studies where subjects are selected based on their disease status, whereas their exposure history (covariate values) is determined retrospectively. We used a *balanced* sample, where we retained all fire occurrences and selected a simple random sample of the same size from the controls; this procedure was repeated to create 500 balanced samples. Doing so induces a deterministic offset intercept term into the logistic modelling framework, $\log(p_1/p_0)$, in the linear predictor $g(.)$ i.e. (for details, see section 6.3 of Hosmer and Lemeshow 2000:

$$g(\pi_i^*) = \log(p_1/p_0) + \boldsymbol{\beta}^t \underline{\boldsymbol{x}}_i.$$

where $\pi_i^* = P(Y_i = 1|\underline{\boldsymbol{x}}_i, s_i = 1)$, s_i is selection status (0 or 1) of the ith voxel; p_1 and p_0 are selection probabilities of cases and controls respectively, which can be determined from respective sampling proportions. For our balanced sampling design, $p_1 = 1$ and p_0 is the ratio of the number of controls to the number of cases in the *entire* dataset. We refer the reader to Woolford *et al.* (2011) for a detailed discussion on this topic in the context of

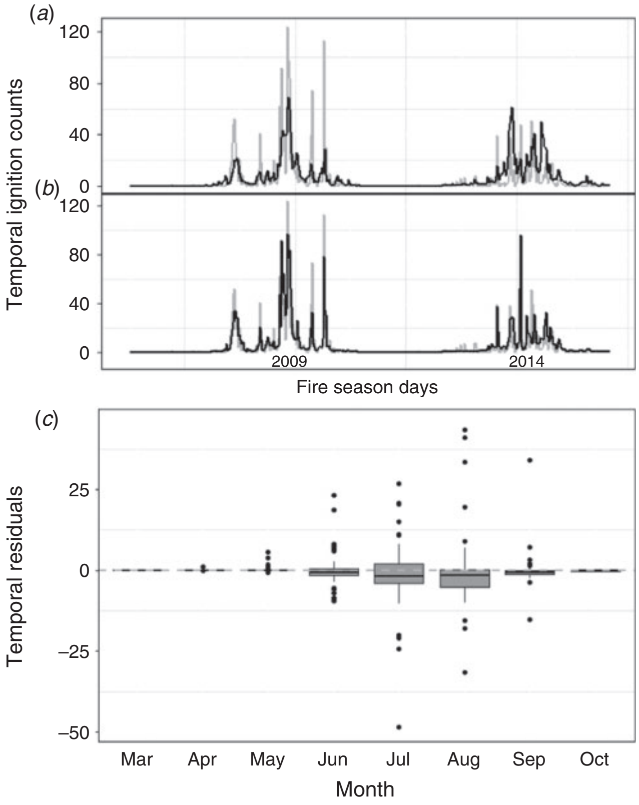

Fig. 3. (*a*, *b*) Daily observed (grey line) and predicted (black line) fire occurrence counts in the 2009 and 2014 test years in British Columbia in the (*a*) Predicted Lightning-Caused Fire (PLCF), and (*b*) Observed Lightning-Caused Fire (OLCF) models. (*c*) Monthly distribution of temporal residual counts for the OLCF model computed over the 2009 and 2014 test years.

forest fire occurrence prediction and its connection to case-control studies. We also comment further on the use of balanced samples in the *Discussion*.

Model fitting

We define two types of fire occurrence counts, temporal and spatial, as follows:

$N^{(H)}_{G,j}$ (temporal counts): total observed human-caused fires over all G grid cells on the j^{th} day; and $N^{(H)}_{i,S}$ (spatial counts): total observed human-caused fires over a given set of S days in cell i. Here, the response variable $Y^{(H)}_{i,j}$ is independently distributed over day–cell combinations, so that counts of occurrences can be predicted as (see Preisler *et al.* 2009 for a similar example):

$$\hat{N}^{(H)}_{G,j} = \sum\nolimits_{i \in G,j} \hat{\pi}^{(H)}_{i,j} \text{ and } \hat{N}^{(H)}_{i,S} = \sum\nolimits_{i,j \in S} \hat{\pi}^{H}_{i,j},$$

where $\hat{\pi}^{(H)}_{i,j}$ is the probability of at least one human-caused fire occurring in voxel (i, j). Similarly, we denote various spatial and temporal count predictors for the two lightning-caused models as $\hat{N}^{(L-PLCF)}_{G,j}$, $\hat{N}^{(L-OLCF)}_{G,j}$, $\hat{N}^{(L-PLCF)}_{i,S}$ and $\hat{N}^{(L-OLCF)}_{i,S}$.

We fit an *ensemble* of 500 individual OLCF, PLCF and HCF models to the corresponding samples to predict day–cell fire occurrence probabilities over grid voxels; then, we determine the average value of the covariate coefficients over all 500 model fits (in our application, each individual model in the ensemble has the same form and covariates). We opted for a large number of datasets in our ensemble ($M = 500$) because: (i) generally, more stable variable rankings are achieved with increasing values of M, and (ii) it allows sufficient coefficient variability in individual fits by exploring the samples space associated with $P(\boldsymbol{X}|Y) = 0$), thereby improving the predictive skill of the ensemble mean. The ensemble probabilities are computed as follows: $\hat{\pi} = g^{-1}\left(\sum\nolimits_{j=1}^{M} g_j(.)/M\right)$, where $g_j(.)$ denotes the estimated logistic link function for the j^{th} model fit. We evaluate predictive skill of the three models (OLCF, PLCF, HCF) based on these ensemble probabilities.

Model evaluation

We split the 34 years of the study dataset (1981–2014; without lightning strikes) into training (1981–2008) and a future test datasets (2009–14) for both HCF and PLCF models. This leaves ~83 and 89% of the total lightning- and human-caused fire occurrence observations respectively for model training and the rest to evaluate predictive skill of the fitted models on future fire seasons. We restrict the OLCF model dataset to 16 years (1999–2014) with consistent cloud-to-ground lightning strike data from the CLDN. We therefore train the OLCF model on 14 years and reserve the 2009 and 2014 fire seasons as test data to evaluate model performance. The 2009 fire season registered an unusually high number of lightning-caused fires over a short timespan (Fig. 3*b*), allowing a starker comparison between the OLCF and PLCF models during surges in fire occurrence.

We evaluate model performance in terms of Receiver Operating Characteristic (ROC) curve analysis (Hosmer and Lemeshow 2000) and root-mean-square prediction error (Table 2), and by visualising prediction bias of count residuals in temporal and spatial dimensions (Figs 3–5). Here, temporal and spatial fire occurrence count residuals for the HCF model are defined as follows:

$$\text{Temporal residuals}: RT^{(H)}_{G,j} = N^{(H)}_{G,j} - \hat{N}^{(H)}_{G,j}$$

$$\text{Spatial residuals}: RS^{(H)}_{i,S_1} = N^{(H)}_{i,S_i} - \hat{N}^{(H)}_{i,S_1},$$

where $N^{(H)}_{i,S_1}$ denotes the number of observed human-caused fires for days S_1 in the i^{th} cell during the fire seasons in the test dataset (2009–14). Similarly, count residuals for the lightning-caused models are notated as $RT^{(L-PLCF)}_{G,j}$, $RT^{(L-OLCF)}_{G,j}$, $RS^{(L-PLCF)}_{i,S_1}$, $RS^{(L-OLCF)}_{i,S_2}$ and $RS^{(L-PLCF)}_{i,S_2}$; where S_2 is the set of days during the 2009 and 2014 fire seasons that comprise the test dataset for the OLCF model. We also report relative root-mean-square prediction error (RRMSPE) based on these residual counts, computed as follows (Table 2):

$$RRMSPE = \sqrt{m^{-1} \sum\nolimits_{k=1}^{m} \left[\left(\theta_k - \hat{\theta}_k\right) / \left(\theta_k + 1\right)\right]^2},$$

Table 2. Receiver operating characteristic (ROC) analysis for predicting individual ignitions (area under curve, AUC, sensitivity and specificity) and relative root-mean-square prediction error (RRMSPE) for predicting number of fire occurrence counts in the Observed Lightning-Caused Fire (OLCF), Predicted Lightning-Caused Fire (PLCF) and Human-Caused Fire (HCF) fire models

Years in parentheses correspond to fire seasons in the test datasets

Model	ROC analysis			RRMSPE	
	AUC	Sensitivity	Specificity	Temporal	Spatial
OLCF (2009, 2014)	0.955	0.905	0.861	1.0487	0.5456
PLCF (2009, 2014)	0.924	0.905	0.790	2.1980	0.5859
PLCF (2009–14)	0.929	0.898	0.814	2.7537	0.6102
HCF (2009–14)	0.913	0.866	0.811	1.1001	0.7702

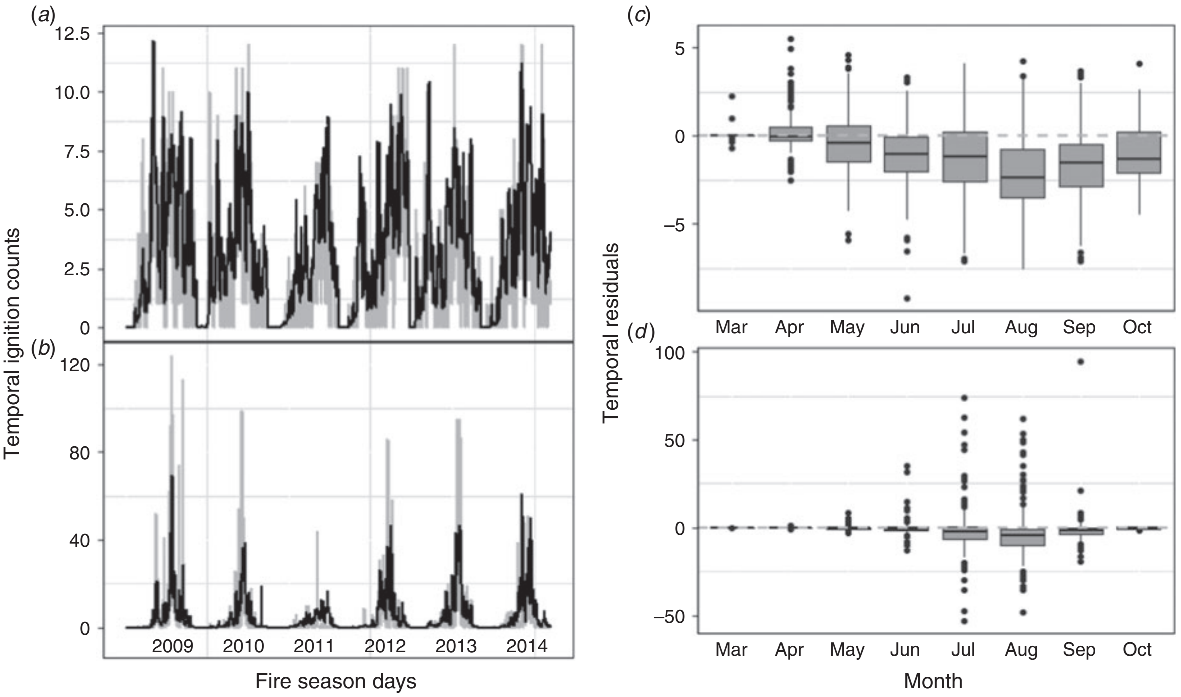

Fig. 4. (*a*, *b*) Daily observed (grey line) and predicted (black line) fire occurrence counts across six test years (2009–14) in BC in the (*a*) Human-Caused Fire (HCF), and (*b*) Predicted Lightning-Caused Fire (PLCF) models. (*c, d*) Monthly distribution of temporal residual counts corresponding to models on the left computed over the 2009–14 test years.

where m is the number of predictions involved and θ_k is a fire occurrence count.

We also compute prediction sensitivity and specificity (Table 2), where the threshold for classifying fire occurrence probabilities as 0 (no fires) or 1 (at least one fire) was based on the Youden index criterion associated with the computed ROC curves (Youden 1950; Hand 2012).

Variable ranking

In this study, we employ a novel covariate ranking algorithm (summarised in Supplementary material) that exploits two key features of our analysis methodology: (i) repeated sampling from controls to create a large number of balanced datasets; (ii) automatic variable selection performed by lasso-logistic when fitted to a single balanced dataset. Consequently, our ranking algorithm involves two major steps: (1) we independently fit the lasso-logistic model to $M = 500$ balanced datasets created via case-control sampling. (2) We average the *standardised* regression coefficients in the M model fits to derive rank metrics for each of the P covariates in $\boldsymbol{x}$. In order to determine a reduced set of the most influential covariates, we further compute an index p_{Drop_i} – the proportion of times the i^{th} covariate is dropped from M lasso-logistic model fits – as follows: $p_{Drop_i} = \sum_{j=1}^{M} I_{\beta_{i,j}=0}/M$.

Results

We fitted three models: OLCF, PLCF and HCF for different fire management applications, as described in the *Objectives* subsection. In this section, we contrast the predictive performance and variable importance among the three models, and evaluate the skill of the ensemble modelling approach.

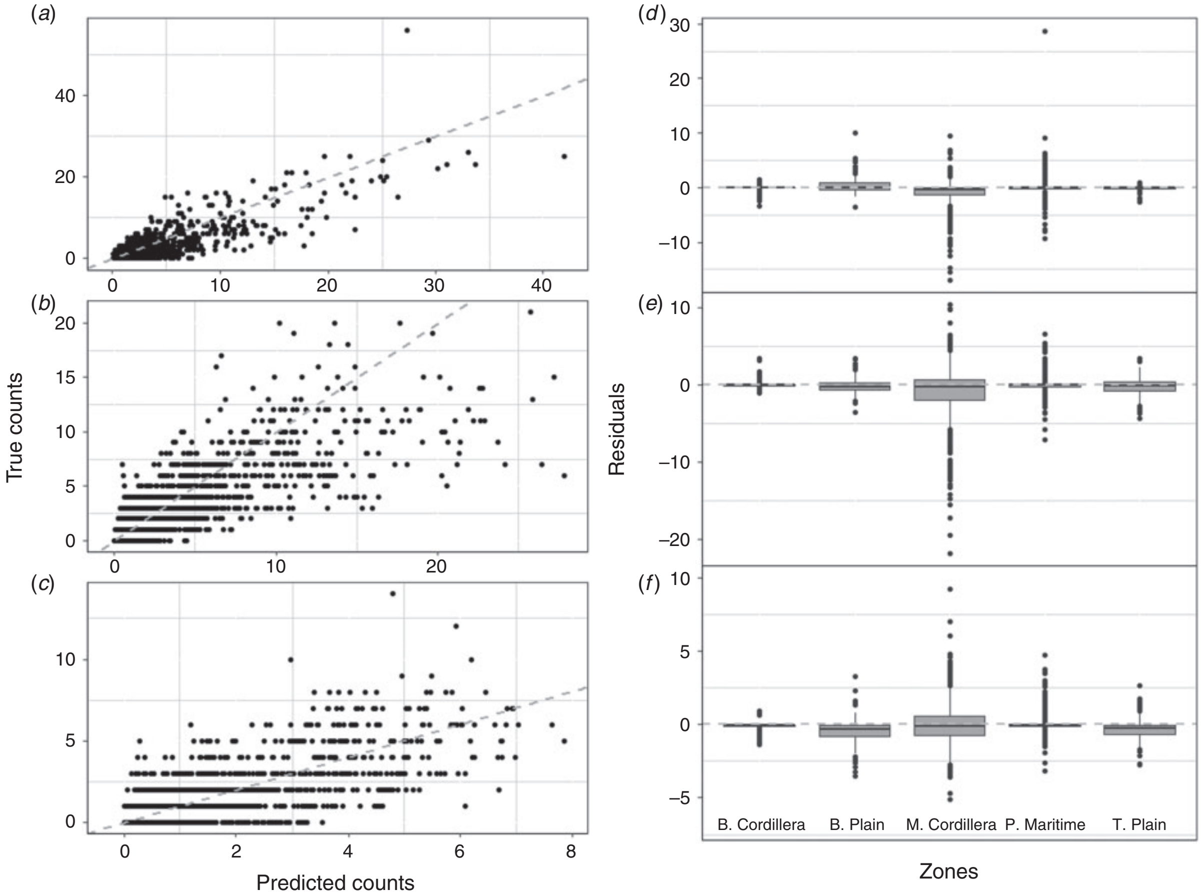

Fig. 5. (*a–c*) Observed *v.* predicted spatial ignition counts for 2541 grid cells computed over the test years. (*a*) Human-Caused Fire (HCF), and (*b*) Predicted Lightning-Caused Fire (PLCF) models with counts computed over 2009–14 test years; and (*c*) Observed Lightning-Caused Fire (OLCF) model with counts computed over 2009 and 2014. (*d–f*) Distribution of spatial observed *v.* predicted fire occurrence count residuals grouped by ecozone (see Fig. 1*b*) corresponding to the models and years on the left.

Predictive performance

We examined the predictive skill of the models in regard to ROC characteristics, provincial-scale time series, and temporal and spatial residuals. The accuracy of all the models in predicting individual fire occurrences for a given day–cell combination on test data was very good (area under receiver-operator curve (AUC) > 0.9); AUC, sensitivity and specificity increased in the order of HCF < PLCF < OLCF (Table 2). Sensitivity was greater than specificity for all models; however, slight to modest overprediction of fire occurrences (and the cost of being overprepared) is more desirable than underprediction in fire and emergency management. The increase in skill due to having information on ignition sources (e.g. lightning strike location and timing) is evident from a large jump in specificity (fewer false positives) from 0.790 and 0.811 for the PLCF and HCF models, to 0.861 for the OLCF model. However, the specificity under OLCF is still below 0.9, indicating that the model can often predict a fire occurrence following lightning strikes that failed to result in an actual reported fire.

In order to evaluate the performance of the models in predicting daily fire occurrences for provincial-scale preparedness planning, we compared the predicted (sum of the cell-based probability values over all cells on a particular day) *v.* the observed daily fire counts in the test years. Not surprisingly, the OLCF model was better than the PLCF model in capturing the peaks in daily fire occurrences in 2009 and 2014 that are due to surges in lightning strikes over broad geographic scales. This is also evident from the residual plot (of under- and overpredictions) in Fig. 3*b*, and the RRMSPE values of 1.049 and 2.198 for the OLCF and PLCF models respectively (Table 2). The daily observed *v.* PLCF-model-predicted fire occurrences are shown again in Fig. 4*b*, but in contrast to the HCF model (Fig. 4*a*), and for the 2009–14 fire seasons. Both of these models track intraseasonal fluctuations in fire occurrence quite well. The PLCF model indicates the occurrence of spikes (approximately >40 fire occurrences per day) in lightning-caused fires, but underestimates their magnitude in all years but 2014 (Fig. 4*b*). Otherwise, temporal residual counts for the PLCF model (Fig. 4-days) are generally unbiased, showing no intraseasonal trend (RMSPE 2.7537; Table 2). The peaks in daily human-caused fire occurrence are smaller (approximately >10 fire occurrences per day) than for lightning fires in BC, and are captured reasonably well by the HCF model (Fig. 4*a*), although the temporal residuals (Fig. 4*c*) show some downward

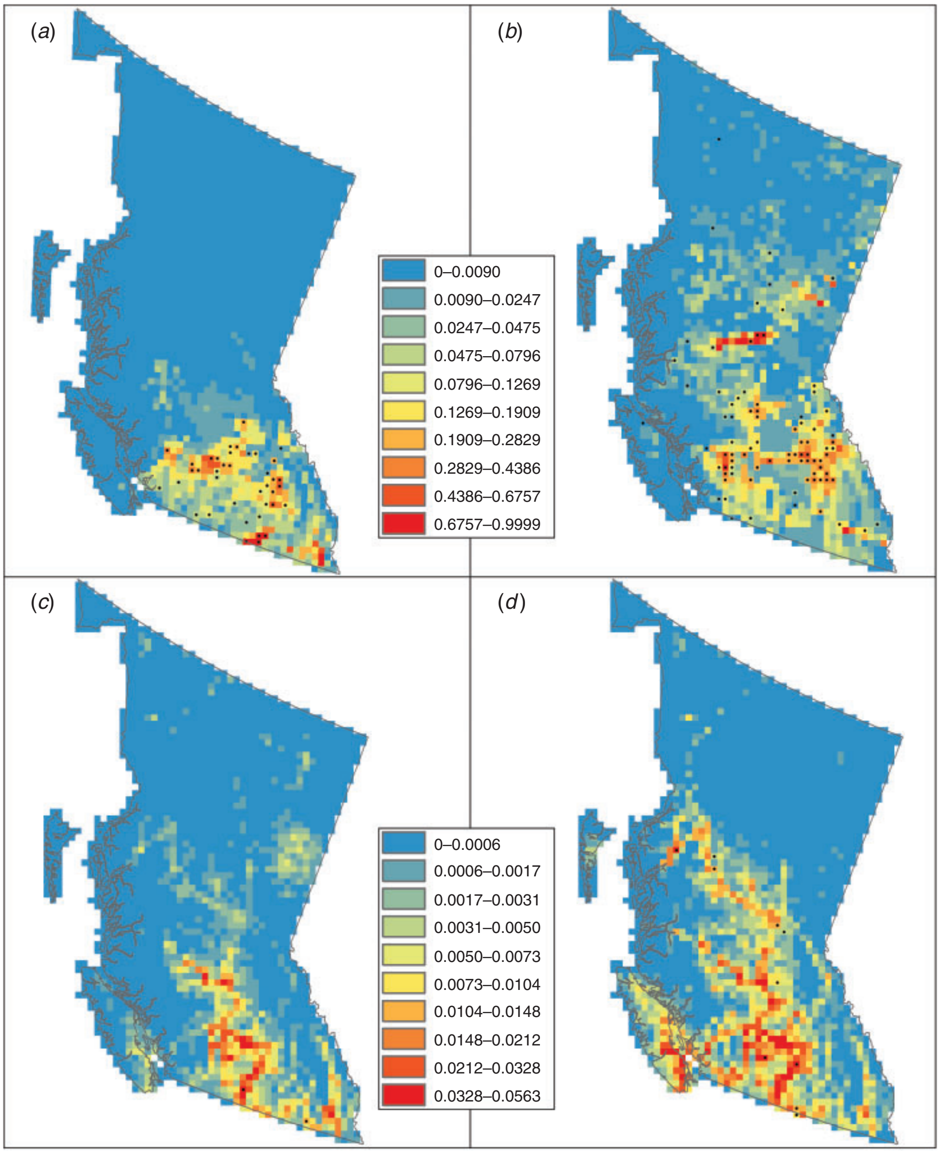

Fig. 6. Predicted probability of a fire occurring within a 400-km^2 cell and locations of observed fires (●) on 4 days in the test dataset. (*a*) Lightning-caused fires on 27 July 2009: predicted 49.3, observed 39; (*b*) lightning-caused fires on 1 August 2009: predicted 84.4, observed 84. (*c*) Person-caused fires on 10 August 2009: predicted 3.8, observed 2; (*d*) person-caused fires on 3 August 2009: predicted 7.4, observed 10.

bias (overprediction) starting from June onwards (RMSPE 1.1001; Table 2). This is mainly driven by the model's false positive rate (Table 2, specificity), which often leads to predicting one or more human-caused fires on days when none occur. Notice, however, that this downward bias is not present in residuals computed for the training dataset (1981–2008; Fig. S1).

The distribution of spatial residuals (from counts of predicted and observed fires in a single grid cell over corresponding test years) is generally unbiased for all models across ecozones, with most residuals tightly concentrated around zero (Fig. 5*d–f*), although there is a greater range of residuals in the Montane Cordillera ecozone where most fires occur. Spatial RMSPE scores increase (accuracy decreases) in the series $0.546 < 0.586 < 0.7702$ for the OLCF, PLCF and HCF models respectively (Table 2), indicating that it is more challenging to predict where human-caused fires may occur. Spatial residuals in the PLCF model are biased downwards (overprediction) for grid cells that accumulate a higher number of fires over multiple seasons. (Fig. 5*b*), possibly owing to the influence of the baseline lightning risk covariate. Fig. 6 provides an example of predicted *v.* observed values for 4 days in the test dataset.

Ensemble performance

We evaluated the ensemble modelling approach in two ways. First, we computed the distribution of AUC values over the 500 individual model fits. Although the AUC values for the ensemble models (Table 2) were approximately one standard

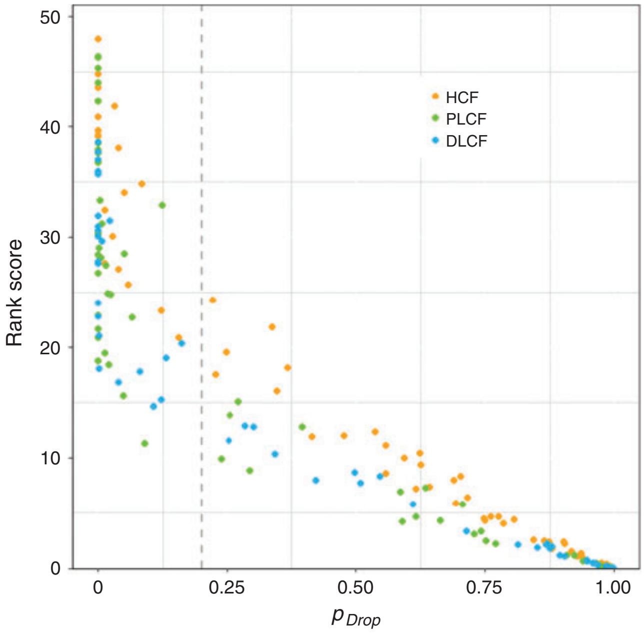

Fig. 7. Relationship between p_{Drop} and rank score, where each point on the plot represents a covariate in the three models. Dashed vertical line shows that there is a large change in p_{Drop} around the 0.2 threshold. Variables above this threshold (23, 29 and 24) are the most influential predictors in the Observed Lightning-Caused Fire (OLCF), Predicted Lightning-Caused Fire (PLCF) and Human-Caused Fire (HCF) models, respectively.

deviation higher than the average of 500 value, we view this as an advantage over using a single model. Second, we fitted three additional OLCF, PLCF and HCLF models using OLR (Tables S6–S7) where: (i) only the basic covariates in each model were included, i.e. excluding the baseline risk and derived covariates; (ii) a stepwise regression algorithm was employed for model selection; and (iii) models were fitted to a single balanced dataset in each case. The AUC value from OLR was lower: ~4.5% for the HCF and PLCF models, and 4.7% for the OLCF model, clearly showing that our framework has greater predictive skill than a less involved benchmark approach.

Variable selection and importance

The variable rank score (Supplementary Material 2, Eqn S3), average *standardised* regression coefficient $\breve{\beta}_i^{\Delta}$ and p_{Drop} index for all of the variables that were included in the OLCF, PLCF and HCF models are given in Supplementary Material 3, Tables S1–S3 in rank order. It should be noted that the average non-standardised coefficients, $\bar{\beta}_i^{\Delta}$, (not reported here) are implemented in model prediction applications. Fig. 7 shows that rank score, *Rank*(*X*), and p_{Drop} have negative relationships, with a large jump in the latter at threshold of ~0.2, forming two distinct clusters of covariates. We consider that all covariates in the cluster below 0.2 are the most important predictors in their respective models. We emphasise that p_{Drop} and importance ranks reveal different aspects of how a covariate influences the predictive skill of the ensemble model. Specifically, rank describes a covariate's influence *relative* to all other covariates, whereas p_{Drop} measures how often that covariate is selected (or rejected) in the model. For example, in the PLCF model (Table S2), probability of sustained flaming (PSUF) and ELEVATION^2 have similar p_{Drop} values of 0.014 and 0.012 respectively, but have rank scores of 27.5 and 19.5.

Variable influence

There are 23, 30 and 20 key influential covariates in the OLCF, PLCF and HCF models respectively (Table 3) with $p_{Drop} < 0.2$. It is not of general interest to examine the contribution of each of these variables here; rather, we briefly review the influence of the more important variables in the remainder of this subsection, following the groupings in Table 1.

Baseline risk

Functions of cell-specific baseline fire occurrence rates (lightning risk rank and logit human risk) appear as top predictors in their respective models. Transformations of the baseline rates were used, because use of the original scale resulted in heavy overestimation of counts in the Montane Cordillera ecozone. Average standardised coefficients of logit-human rate and its square are 0.284 and –1.634 (Table S3), revealing a quadratic effect (concave downwards) of logit-human rate on fire occurrence probability that plateaus for higher values of baseline human rate. The baseline rates contribute to predictive skill by accounting for the underlying strong spatial structure in the long-term expectations, which is fairly stable over time. Despite the fact that a baseline rate is a function of response random variable, this temporal stability allows future predictions under the fitted models because the baseline predictor assumes the same values over the spatial grid for both our training and validation (future) dataset. Note that no fires were reported in ~33 and 22% of the cells in BC during 1981–2014 (see Fig. 2). The baseline rates help to reduce the false positive rate by allowing more accurate predictions over the regions with no fire activity. It is also important to note here that all subsequent variable effects are departures from the baseline and should be viewed in that context.

Surface weather

Surface fire weather measures were the second most influential group of covariates that are common to all models. RELATIVE HUMIDITY, PRECIPITATION LAG1, SDMC and PSUF are influential in all models. The counter-intuitive positive influence of RELATIVE HUMIDITY and PRECIPITATION in the lightning models may be because lightning is associated with the passage of storm fronts resulting in precipitation and higher humidity on a broad regional basis. TEMPERATURE and BUI have positive influences in the lightning models. In the HCF model, ISI (a function of FFMC and wind speed) and WIND SPEED have a positive influence. In short, lightning-caused fires are favoured by warm, humid days, and human-caused fires by dry, windy days, conditional in both cases on dry forest floor fuels. The influence of the transformed variables and interaction terms are more difficult to interpret. Because the FWI System codes and indices all increase with decreasing fuel moisture content and some with increasing wind speed, and DMC and particularly DC also increase over the fire season, squared effects may provide a levelling-off effect.

Table 3. Importance rank^{A} of the most influential variables by group (average group rank in parentheses) in the Observed Lightning-Caused Fire model (OCLF, 23 of 53 variables), the Predicted Lightning-Caused Fire model (PLCF, 29 of 62 variables) and the Human-Caused Fire model (HCF, 20 of 72 variables)

Variable names followed by (–) indicate that the sign of the coefficient is negative. Between-model contrasts are shown by typeface: variables influential in a single model (plain type); underlined, influential variables common to all models; italic, influential variables common to the lightning models. Ranks, coefficients and pdrop values for all variables are in Tables S1–S3 and variable definitions are in Table 1

OLCF Variables	Rank^{A}	PLCF Variables	Rank	HCF Variables	Rank
Baseline risk	(1)	**Baseline risk**	(1)	**Baseline risk**	(4)
LIGHTNING RISK RANK	1	*LIGHTNING RISK RANK*	1	LOGIT HUMAN RISK^2 (–)	1
				LOGIT HUMAN RISK	6
Ignition indicator	(8)	**Ignition indicator**	(28)	**Ignition indicator**	(22)
LIGHTNING STRIKES	2	SHOWALTER INDEX (–)	6	WUI DISTANCE^2 (–)	22
LIGHTNING STRIKES INDICATOR	3	C-HAINES INDEX (–)	9		
LIGHTNING LAG2 (–)	20	K INDEX	19		
		500 MB ANOMALY (–)	27		
		500 MB TENDENCY (–)	29		
Surface weather	(11)	**Surface weather**	(14)	**Surface weather**	(10)
SDMC	4	*TEMPERATURE*	2	RELATIVE HUMIDITY (–)	2
TEMPERATURE	5	SDMC	3	ISI	4
RELATIVE HUMIDITY	6	RELATIVE HUMIDITY	4	FFMC..ACCUM PRECIP	5
SDMC^2 (–)	7	*SDMC*^{2} (–)	5	PSUF	7
PSUF	10	*DC..ACCUM PRECIP* (–)	7	ACCUMULATED PRECIPITATION	8
BUI	11	PRECIPITATION	11	DC^2	9
DC.. ACCUM PRECIP	12	*BUI*	12	FWI$^{2)}$ (–)	11
PRECIPITATION LAG1 (–)	15	RELATIVE HUMIDITY^2 (–)	13	SDMC	12
PRECIPITATION	16	DC.. TEMPERATURE (–)	14	WIND SPEED	17
DMC^2 (–)	17	DMC.. ACCUM PRECIP (–)	17	PRECIPITATION LAG1 (–)	19
FFMC..ACCUM PRECIP	18	PRECIPITATION LAG^3 (–)	18		
DC^2 (–)	23	PSUF	20		
		PRECIPITATION LAG^2 (–)	21		
		FWI^2 (–)	23		
		PRECIPITATION LAG1 (–)	24		
		FFMC	25		
Geographic	(16)	**Geographic**	(22)	**Geographic**	(14)
LONGITUDE (–)	8	LATITUDE	10	ELEVATION (–)	10
PACIFIC MARITIME	14	ROUGHNESS	15	ROUGHNESS	15
ROUGHNESS	15	PACIFIC MARITIME	16	PACIFIC MARITIME	14
ELEVATION^2	20	TAIGA PLAIN (–)	26	BOREAL PLAIN (–)	16
BOREAL PLAIN	21	*ELEVATION*^{2}	28		
		BOREAL PLAIN	35		
Vegetation	(16)	**Vegetation**	(20)	**Vegetation**	(16)
VEGETATED (–)	9	VEGETATED (–)	8	VEGETATED (–)	13
AVERAGE NDVI	23	CONIFER COVER	22	AVERAGE NDVI (–)	18
		% CONIFER	31		

[A] Importance rank as per the p_{Drop} criterion depicted in Fig. 7 based on 500 model fits.

Ignition indicators

The ignition indicators variables are unique to each model. Not surprisingly, LIGHTNING STRIKES and the LIGHTNING INDICATOR were very influential covariates in the OLCF model, although the LIGHTNING LAG2 has a negative influence, which is possibly a levelling-off effect. Because of the high specificity of these lightning indicators, only 23 variables were required to reach the p_{Drop} threshold of 0.2 in the OLCF model whereas 32 were required in the PLCF model. Five indexes of atmospheric pressure and stability, most importantly the Showalter and Continuous Haines Indexes, were influential in the PLCF model; including these variables increased AUC for the PLCF model by ~2%.

Among the ecumene covariates specific to the HCF model, WUI DISTANCE^2 had a negative influence, as would be expected. Preliminary analysis revealed there was a significant change (reduction) in human-caused fire occurrence in ~1992 as shown in Fig S2. Thus, a CHANGE POINT variable (0,1) was included to indicate pre- and post-1992 epochs; it was the third ranked predictor in the HCF model. Although important in modelling past fires, the change point is not used in future predictions. Whereas WEEKDAY was not among the most influential variables in the HCF model, it has interesting results. Friday–Monday and Tuesday–Thursday have positive and negative influences respectively, consistent with a weekend effect with persistence as shown in

Fig S3*a*. Tuesday (–) and Sunday (+) were the most influential weekdays.

Geographic

LATITUDE and LONGITUDE have a positive influence (south to north and from east to west respectively) in all models, although latitude and longitude were only highly influential in the PLCF and OLCF models respectively. ELEVATION^2 and ROUGHNESS have a positive influence in the lightning models, consistent with lightning in the mountains. Although most fires occur in the Montane Cordillera ecozone (Figs S3*b*, S4), PACIFIC MARITIME had a positive influence in all models, whereas BOREAL PLAIN had a positive influence in the OLCF and HCF models, and TAIGA PLAIN in the PLCF model. However, we determined that ecozone effects are confounded with baseline risk and ecumene covariates by refitting the models (ensemble of 30 model fits) without these covariates. Dropping baseline risk results in Montane Cordillera being ranked as an important covariate with a positive influence, whereas Taiga Plain and Boreal Plain have a negative influence. The remaining two ecozones are dropped from the ranking process. Dropping both the baseline risk and ecumene covariates from the HCF model results in Boreal Plain and Montane Cordillera having a positive influence and Taiga Plain and Boreal Cordillera having a negative influence.

Vegetation

Vegetation covariates were among the least influential, possibly because of the quite coarse scale of the modelling framework (20 × 20-km cells) relative to the variation in vegetative cover types. The amount of VEGETATED area had a negative influence in all models, which may be associated with a higher proportion of mountains or developed areas. However, CONIFER COVER (Fig. 1*a*) had a positive influence in the PLCF and HCF models. AVERAGE NDVI, a measure of the seasonal trend in vegetation greenness, has a bell-shaped distribution during the fire season over much of BC, increasing from early spring to a summer peak, then declining; it was specifically included as an index of seasonality. AVERAGE NDVI has a positive influence in the lightning-caused fire models, likely because lightning-caused fires have a similar seasonal trend in BC. However, average NDVI has a negative influence in the HCF model, reflecting the peak in human-caused fires in the spring in some regions.

Discussion

Modelling approach

We used sampling to create *M* balanced datasets that we then used to create an ensemble of *M* fitted models. The advantage of the ensemble is that, although there is the possibility of a single model fit outperforming the ensemble, there is also the strong possibility that it could perform worse. Our problem involved more than 75 covariates of various types (categorical and continuous), including interaction terms (Table 1). We prefer the lasso-logistic model for this problem, because, as compared with other forms of penalised regression forms, e.g. ridge regression with L_2-penalty, the L_1-penalty in lasso allows automatic variable selection by shrinking those coefficients to identically zero that do not improve the model's predictive skill. Furthermore, the covariates with non-zero coefficient estimates are generally readily interpretable. As our focus is on developing models that have superior skill in predicting future fire occurrences for management purposes, optimisation of Eqn 2 using *k*-fold cross-validation allows better generalisation of the model to future fire-weather conditions. Nonetheless, non-linear responses are difficult to estimate. Evaluation of and comparison with machine learning-based models is ongoing and will be reported separately.

We used balanced datasets in our sampling procedure, which is a recommended option in classification problems for rare events data (see, for example, the Balanced Random Forests algorithm in Chen *et al.* 2004). Although it is possible to adjust for an arbitrary choice of case-control sampling ratio by adding an offset term, evidence from literature suggests that classifiers can exhibit substantial variability in predictive accuracy over a broader range of ratios. This effect is especially pronounced for highly imbalanced datasets (see, for instance, Byon *et al.* 2010). That is, sampling ratio can act as an extra tuning parameter that must be learned from training data, for instance via the cross-validation approach. This is an interesting area for investigation of logistic regression models with response-based sampling. This was explored by Woolford *et al.* (2011) who demonstrated that the estimated partial effects in a logistic generalised additive model for person-caused fire occurrence prediction are sensitive to large reductions in the inclusion probability for the non-fire voxels.

Predictive performance

Although the OLCF model has greater predictive skill than the PLCF and HCF models, it is a retrospective model and its application is largely in nowcasting (lightning-caused ignitions that have occurred but may or may not have yet been detected) whereas the latter models have a purely forecast application. Although nowcasting is useful for directing detection efforts following a lightning storm, forecast models have more value for preparedness planning.

As was noted earlier, sharp peaks in the number of lightning-caused fires are a critical process that has very significant impacts on the fire management system and can result in large areas being burned if several fires escape initial attack in high fire danger periods. Although the OLCF model identifies peaks in lightning fire occurrence (approximately >50 fires per day), accurately predicting the number of fires on peak fire days is difficult.

We used a 20 × 20-km grid cell as the spatial sampling unit as a compromise between the resolution of covariates and class imbalance, both of which increase with decreasing cell size. We consider it to be a mesoscale model in the meteorological sense, recognising that there are microscale influences of vegetation and topography on fuel moisture and so on ignition probability (e.g. lightning ignitions will smoulder longer in deep forest floor organic layers; Latham and Williams 2001; Wotton *et al.* 2005) and such environmental detail is not represented at this scale. Finer-scale models could be explored using the sampling methodology introduced here, although there are issues with both the spatial accuracy of historical fire records and incomplete knowledge of vegetation change over time.

Variable selection and importance

Predictive analytics are ultimately limited by our knowledge of the underlying processes and ability to accurately quantify influential factors. In simple models of ignition probability such as the 2-min test fires reanalysed by Beverly and Wotton (2007), the covariates have very straightforward interpretations. However, the relatively large numbers of parameters in more complex spatiotemporal point process fire occurrence models can make model estimation, interpretation and improvement challenging (Schoenberg 2016). The development of a variable ranking methodology was an important step because further model improvements should result from identifying and enhancing the representation of the most influential processes and covariates.

Because the focus of the present paper was on prediction, we included baseline risk covariates as they were very important in increasing model sensitivity and specificity, recognising that the baseline risk may affect the interpretation of other variables. Transforms of baseline risk were highly ranked in all models. These rates can be viewed as estimates of spatial random effects (one effect for each grid cell) – in principle, a model with 2541 fixed cell-effects – but fitting such a model is practically implausible and is not done here. By calculating these baseline values and using them as a predictor, we are estimating the relationship between fire occurrence risk and the baseline, rather than having a separate fixed effect for each individual cell. This baseline risk effect accounts for otherwise unspecified cell-specific factors or interactions related to the intensity of human and lightning ignitions or the ignitability of fuels not well represented by other covariates. For example, road density was influential in the simple OLR HCF model without baseline risk (Table S6) but it alone is an incomplete representation of the intensity of human activity in models.

There are alternative methods for representing baseline risks. For example, Brillinger *et al.* (2003) modelled baseline fire occurrence risk using spatial and seasonal smoothers. A similar approach was employed by the models of both fire occurrence and large fire risk of Preisler *et al.* (2004). An advantage of the use of smoothers is that they can account for non-linear effects. As such, further refinement of our methodology might include temporal variation in baseline risk. We tested monthly (*v.* annual) human-caused baseline fire risk but it did not improve HCF model fit. This may be an effect of including seasonally changing variables, as was also noted by Preisler *et al.* (2004). Furthermore, human activity and its relationships with fire occurrence risk may not be stationary at any location over the 30-year period owing to changing land use; baseline risk may be sharpened, perhaps by using a weighted moving average of annual events.

Weather and FWI System values, their transformations and derivatives were consistently important covariates in all models in as much as they represent fuel moisture and ignitability. It is noteworthy that SDMC and PSUF (Lawson and Armitage 1997, probability of sustained ignition eqn 9D), which were used in a lightning fire occurrence model for Ontario (Wotton and Martell 2005), were also influential in the BC models. Several PSUF models have been developed through experimentation in different vegetation types (Beverly and Wotton 2007) that could be evaluated for different ecological conditions.

The OLCF model is conditional on the known time and location of lightning strikes, and this likely explains the high AUC. In contrast, the PLCF and HCF models estimate the outcome of two processes – the likely location of lightning strikes and anthropogenic ignition sources, and the occurrence probability. The probable locations of lightning and human fire-causing activity are but crudely estimated through the use of a few atmospheric stability indices in the case of lightning, and ecumene metrics such as population and road length for human activity; the latter variables have no daily or seasonal dynamics that mirror human activity. Further model development might separate the intensity and occurrence probability processes by developing more sophisticated models or indices of lightning fire potential, or temporally varying indices of human activity in wildlands. For example, information from anonymised mobile phone geolocation data may be a promising source of high-resolution spatiotemporal information on human presence in wildlands with cell coverage (e.g. Deville *et al.* 2014).

Whether lightning is accompanied by rain is critical to whether ignitions are sustained but is difficult to estimate in fairly sparse weather station networks. Precipitation distribution varies within thunderstorm cells, and because there is a higher probability of fire starts on the edges of cells, the distance to storm centres might provide a proxy (Woolford and Braun 2007).

An illustration of the spatial pattern of predicted and observed fires, and of the total number of predicted and observed fires across British Columbia for 4 peak days in 2009 is shown in Fig. 6. We anticipate that such spatial occurrence maps and corresponding estimates of new fire starts will be a valuable tool for prevention and preparedness planning and decision-making, including public advisories and access restrictions, resource sharing between fire control agencies, man-up or readiness levels, prepositioning fire crews and aircraft, and detection aircraft routing. The models developed here, as well as similar models for other regions of Canada, are being implemented in the CWFIS. Data-driven modelling requires good data; ongoing collaboration with fire managers will be important in improving data quality.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

The cooperation of Canadian provincial and territorial fire management agencies and the Desert Research Institute in providing fire weather data is gratefully acknowledged. R. Benton, G. Thandi and K. Omendje assisted with data compilation. S. Magnussen and M. Wotton advised on study design and variable selection. This project was supported by the Department of National Defence Canadian Safety and Security Program. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) (funding reference nos RGPIN-2015–04221 and RGPIN-2014–06187) and of a Canadian Statistical Sciences Institute (CANSSI) Collaborative Research Team grant.

References

Alexander GW (1927) Lightning storms and forest fires in the state of Washington. *Monthly Weather Review* **55**, 122–129. doi:10.1175/1520-0493(1927)55<122:LSAFFI>2.0.CO;2

Beall HW (1934) A practical test of the accuracy of forest-fire hazard charts. *Forestry Chronicle* **10**, 56–65. doi:10.5558/TFC10056-1

Beaudoin A, Bernier PY, Guindon L, Villemaire P, Guo XJ, Stinson G, Bergeron T, Magnussen S, Hall RJ (2014) Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. *Canadian Journal of Forest Research* **44**, 521–532. doi:10.1139/CJFR-2013-0401

Beverly JL, Wotton BM (2007) Modelling the probability of sustained flaming: predictive value of fire weather index components compared with observations of site weather and fuel moisture conditions. *International Journal of Wildland Fire* **16**, 161–173. doi:10.1071/WF06072

Brillinger DR, Preisler HK, Benoit JW (2003) Risk assessment: a forest fire example. In 'Statistics and science: a Festschrift for Terry Speed'. pp. 177–196, Vol. 40 of IMS Lecture Notes Monograph Series, DR Goldstein, (ed), (Institute of Mathematical Statistics: Beachwood, OH, USA). doi:10.1214/LNMS/1215091142.. doi:10.1214/LNMS/1215091142

Bruce D (1963) How many fires? *Fire Control Notes* **24**, 45–51.

Burrows WR, King P, Lewis PJ, Kochtubajda B, Snyder B, Turcotte V (2002) Lightning occurrence patterns over Canada and adjacent United States from lightning detection network observations. *Atmosphere-ocean* **40**, 59–80. doi:10.3137/AO.400104

Byon E, Shrivastava AK, Ding Y (2010) A classification procedure for highly imbalanced class sizes. *IIE Transactions* **42**, 288–303. doi:10.1080/07408170903228967

Camp PE, Krawchuk MA (2017) Spatially varying constraints of human-caused fire occurrence in BC, Canada. *International Journal of Wildland Fire* **26**, 219–229. doi:10.1071/WF16108

Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. Technical report #666, Department of Statistics, University of California. (Berkeley, CA, USA).

Costafreda-Aumedes S, Comas C, Vega-Garcia C (2017) Human-caused fire occurrence modelling in perspective: a review. *International Journal of Wildland Fire* **26**, 983–998. doi:10.1071/WF17026

Crosby JS (1954) Probability of fire occurrence can be predicted. USDA Forest Service, Central States Forest Experiment Station, Technical Paper 143. (Columbus, OH, USA)

Cunningham AA, Martell DL (1973) A stochastic model for the occurrence of man-caused forest fires. *Canadian Journal of Forest Research* **3**, 282–287. doi:10.1139/X73-038

De Smith MJ, Goodchild MF, Longley PA (2015) 'Geospatial analysis', 5th edn. (Matador: Leicester, UK)

Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan A, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 15888–15893. doi:10.1073/PNAS.1408439111

Dockendorff D, Spring K (2005) The Canadian Lightning Detection Network – Novel approaches for performance measurement and network management. In 'WMO technical conference on instruments and methods of observation (TECO-2005)', 4–7 May 2005, Bucharest, Romania. pp. 62–67, Instruments and Observing Methods, Report No. 82. (World Meteorological Organization: Geneva, Switzerland)

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22. doi:10.18637/JSS.V033.I01

Gilbert DE, Zala C (1987) A reliability study of the lightning-locating network in BC. *Canadian Journal of Forest Research* **17**, 1060–1065. doi:10.1139/X87-162

Gillis MD, Omule AY, Brierley T (2005) Monitoring Canada's forests: the National Forest Inventory. *Forestry Chronicle* **81**, 214–221. doi:10.5558/TFC81214-2

Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications* **73**, 220–239. doi:10.1016/J.ESWA.2016.12.035

Hand DJ (2012) Assessing the performance of classification methods. *International Statistical Review* **80**, 400–414. doi:10.1111/J.1751-5823.2012.00183.X

Hardy CC, Hardy CE (2007) Fire danger rating in the United States of America: an evolution since 1916. *International Journal of Wildland Fire* **16**, 217–231. doi:10.1071/WF06076

Hornby LG (1936) Fire control planning in the northern Rocky Mountain region. USDA Forest Service, Northern Rocky Mountain Forest and Range Experiment Station. (Missoula, MT, USA).

Hosmer DW, Lemeshow S (2000) 'Applied logistic regression', 2nd edn. (Wiley: New York, NY, USA).

Jackson AW (1968) Weather forecasting applied to forest fire protection in BC. Canadian Department of Transportation Meteorological Branch, Technical Memorandum Tec-693. (Ottawa, ON, Canada).

Janz B, Nimchuk N (1985) The 500 mb anomaly chart: a useful fire management tool. In '8th conference on fire and forest meteorology', 29 April–2 May, 1985, Detroit, MI. (Eds LR Donoghue, RE Martin) pp. 233–238. (Society of American Foresters: Bethesda, MD, USA).

Johnston LM, Flannigan MD (2018) Mapping Canadian wildland fire interface areas. *International Journal of Wildland Fire* **27**, 1–14. doi:10.1071/WF16221

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**, 437–472. doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Latham D, Williams E (2001) Lightning and forest fires. In 'Forest fires'. (Eds EA Johnson, K Miyanishi) pp. 375–418. (Academic Press: San Diego, CA, USA).

Lawson BD, Armitage OB (1997) Ignition probability equations for some Canadian fuel types. Report to the Canadian Committee on Forest Fire Management. Ember Research Services Ltd. (Victoria, BC, Canada).

Lawson BD, Armitage OB (2008) Weather guide for the Canadian Forest Fire Danger Rating System. Canadian Forest Service, Northern Forestry Centre. (Edmonton, AB, Canada).

Lee BS, Alexander ME, Hawkes BC, Lynham TJ, Stocks BJ, Englefield P (2002) Information systems in support of wildland fire management decision making in Canada. *Computers and Electronics in Agriculture* **37**, 185–198. doi:10.1016/S0168-1699(02)00120-5

Lee S, Lee H, Abeel P, Ng A (2006) Efficient L1 – regularized logistic regression. In 'Proceedings of the 21st national conference on artificial intelligence (AAAI-06)', 16–20 July, Boston, MA. AAAI-06, pp. 401–408. A Cohn (ed), (American Association for Artificial Intelligence Press, Palo Alto, CA).

Magnussen S, Taylor SW (2012) Prediction of daily lightning-and human-caused fires in BC. *International Journal of Wildland Fire* **21**, 342–356. doi:10.1071/WF11088

Melrose GP, Holmgren W (1932) Lightning and forest fires in the southern interior region of BC. *Forestry Chronicle* **8**, 158–170. doi:10.5558/TFC8158-3

Mesinger F, DiMego G, Kalnay E, Mitchell K (2006) North American regional reanalysis. *Bulletin of the American Meteorological Society* **87**, 343–360. doi:10.1175/BAMS-87-3-343

Meyn A, Schmidtlein S, Taylor SW, Girardin MP, Thonicke K, Cramer W (2010) Spatial variation of trends in wildfire and summer drought in BC, Canada, 1920–2000. *International Journal of Wildland Fire* **19**, 272–283. doi:10.1071/WF09055

Mills GA, McCaw L (2010) Atmospheric stability environments and fire weather in Australia – extending the Haines Index. CAWCR Technical Report No. 20. (CSIRO and the Bureau of Meteorology, Melbourne, Vic., Australia).

Moore RD, Spittlehouse DL, Whitfield PH, Stahl K (2010) Weather and climate. In 'Compendium of forest hydrology and geomorphology in BC'. (Eds RG Pike, TE Redding, RD Moore, RD Winkler, KD Bladon) pp. 47–84. (BC Ministry of Forests and Range, Research Branch and FORREX Forest Research Extension Partnership: Victoria and Kamloops, BC, Canada).

Nimchuk N (1983) Wildfire behavior associated with upper ridge breakdown. Alberta Ministry of Energy and Natural Resources, Forest Service, ENR Report T/50. (Edmonton, AB, Canada).

Noble DV (1926) Relative humidity and the incidence of forest fires. *American Meteorological Society Bulletin* **7**, 74–77.

Noggle RC, Krider EP, Vance DL, Barker KB (1976) A lightning direction finding system for forest fire detection. In 'The 4th national conference fire and forest meteorology', 16–18 November 1976, St Louis, MO. (Eds DH Baker, MA Fosberg) USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, General Technical Report RM-32, pp. 31–36. (Fort Collins, CO, USA).

Nychka D, Furrer R, Paige J, Sain S (2017) R package 'fields' version 9.6: Tools for spatial data. Available at https://www.rdocumentation.org/packages/fields/versions/9.6 (Verified 12 November 2019). doi:10.5065/D6W957CT

Preisler HK, Westerling AL (2007) Statistical model for forecasting monthly large wildfire events in western United States. *Journal of Applied Meteorology and Climatology* **46**, 1020–1030. doi:10.1175/JAM2513.1

Preisler HK, Brillinger DR, Burgan RE, Benoit JW (2004) Probability-based models for estimation of wildfire risk. *International Journal of Wildland Fire* **13**, 133–142. doi:10.1071/WF02061

Preisler HK, Burgan RE, Eidenshink JC, Klaver JM, Klaver RW (2009) Forecasting distributions of large federal-lands fires utilizing satellite and gridded weather information. *International Journal of Wildland Fire* **18**, 508–516. doi:10.1071/WF08032

Saari E (1923) Kuloista etupäässä Suomen vationmetsiä silmällä pitäen (Forest fires in Finland with special reference to the state forest. Statistical investigation). *Acta Forestalia Fennica* **26**, 1–143.

Schoenberg FP (2016) A note on the consistent estimation of spatial-temporal point process parameters. *Statistica Sinica* **2**, 861–879. doi:10.5705/SS.2014.150

Show SB, Kotok EI (1923) Forest fires in California 1911–1920 – An analytical study. Department Circular 243. USDA. (Washington, DC, USA).

Simard AJ, Valenzuela J (1972) A climatological summary of the Canadian Forest Fire Weather Index. Canada Department of the Environment, Canadian Forestry Service, Forest Fire Research Institute, Information Report Number FF-X-34. (Ottawa, ON, Canada).

Stahl K, Moore RD, McKendry IG (2006) The role of synoptic-scale circulation in the linkage between large-scale ocean atmosphere indices and winter surface climate in British Columbia, Canada. *International Journal of Climatology* **26**, 541–560. doi:10.1002/JOC.1268

Statistics Canada (2012) GeoSuite, reference guide. Census year 2011. Ministry of Industry, Statistics Canada, Catalogue no. 92–150-G. (Ottawa, ON, Canada).

Statistics Canada (2015) Road network file, reference guide. Ministry of Industry, Statistics Canada, Catalogue no. 92–500-G. (Ottawa, ON, Canada).

Stull R (2015) 'Practical meteorology: an algebra-based survey of atmospheric science.' (University of BC: Vancouver, BC, Canada).

Taylor SW, Alexander ME (2006) Science, technology, and human factors in fire danger rating: the Canadian experience. *International Journal of Wildland Fire* **15**, 121–135. doi:10.1071/WF05021

Taylor SW, Woolford DG, Dean CB, Martell DL (2013) Wildfire prediction to inform management: statistical science challenges. *Statistical Science* **28**, 586–615. doi:10.1214/13-STS451

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological* **58**, 267–288. doi:10.1111/J.2517-6161.1996.TB02080.X

Toth Z, Desmarais J-G, Brunet G, Zhu Y, Verret R, Wobus R, Hogue R, Cui B (2005) The North American Ensemble Forecast System (NAEFS). *Geophysical Research Abstracts* **7**, 02501.

Van Wagner CE (1987) Development and structure of the Canadian Forest Fire Weather Index System. Canadian Forestry Service, Forest Technology Report 35. (Ottawa, ON, Canada)

Vermote E, Justice C, Csiszar I, Eidenshink J, Myneni R, Baret F, Masuoka E, Wolfe R, Claverie M, NOAA CDR Program (2014) NOAA Climate Data Record (CDR) of Normalized Difference Vegetation Index (NDVI), Version 4. (NOAA National Climatic Data Center). Available at https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00813# (Verified 12 November 2019). . doi:10.7289/V5PZ56R6

Vilar L, Woolford DG, Martell DL, Martn MP (2010) A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. *International Journal of Wildland Fire* **19**, 325–337.

Wang H, Boissel J-P, Nony P (2009) Revisiting the relationship between baseline risk and risk under treatment. *Emerging Themes in Epidemiology* **6**, Article 1. doi:10.1186/1742-7622-6-1

Wang X, Wotton BM, Cantin AS, Parisien MA, Anderson K, Moore B, Flannigan MD (2017) cffdrs: an *R* package for the Canadian Forest Fire Danger Rating System. *Ecological Processes* **6**, Article 5. doi:10.1186/S13717-017-0070-Z

Wiken EB (1986) 'Terrestrial ecozones of Canada.' (Environment Canada, Lands Directorate: Ottawa, ON, Canada)

Woolford DG, Braun WJ (2007) Convergent data sharpening for the identification and tracking of spatial temporal centers of lightning activity. *Environmetrics* **18**, 461–479. doi:10.1002/ENV.815

Woolford DG, Bellhouse DR, Braun WJ, Dean CB, Martell DL, Sun J (2011) A spatiotemporal model for people-caused forest fire occurrence in the Romeo Malette forest. *Journal of Environmental Statistics* **2**, 2–16.

Wotton BM (2009) Interpreting and using outputs from the Canadian Forest Fire Danger Rating System in research applications. *Environmental and Ecological Statistics* **16**, 107–131. doi:10.1007/S10651-007-0084-2

Wotton BM, Flannigan MD (1993) Length of the fire season in a changing climate. *Forestry Chronicle* **69**, 187–192. doi:10.5558/TFC69187-2

Wotton BM, Martell DL (2005) A lightning fire occurrence model for Ontario. *Canadian Journal of Forest Research* **35**, 1389–1401. doi:10.1139/X05-071

Wotton BM, Stocks BJ, Martell DL (2005) An index for tracking sheltered forest floor moisture within the Canadian Forest Fire Weather Index System. *International Journal of Wildland Fire* **14**, 169–182. doi:10.1071/WF04038

Xi DX, Taylor SW, Woolford DG, Dean CB (2019) Statistical models of key components of wildfire risk. *Annual Review of Statistics and its Applications* **6**, 197–222. doi:10.1146/ANNUREV-STATISTICS-031017-100450

Youden WJ (1950) Index for rating diagnostic tests. *Cancer* **3**, 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3