

Risk Management

Random Forests for Wildfire Insurance Applications

Roba Bairakdar^a, Mathieu Boudreault^{1b}, Melina Mailhot^{2c}

¹ Department of Mathematics, University of Quebec at Montreal, ² Department of Mathematics and Statistics, Concordia University - Canada

Keywords: Wildfires, Random forest, Pure premium, Burn probabilities

Variance

Vol. 16, Issue 2, 2023

Homeowners' insurance in wildfire-prone areas can be a very risky business that some insurers may not be willing to undertake. We create an actuarial spatial model for the likelihood of wildfire occurrence over a fine grid map of North America. Several models are used, such as generalized linear models and tree-based machine learning algorithms. A detailed analysis and comparison of the models show a best fit using random forests. Sensitivity tests help in assessing the effect of future changes in the covariates of the model. A downscaling exercise is performed, focusing on some high-risk states and provinces. The model provides the foundation for actuaries to price, reserve, and manage the financial risk from severe wildfires.

Address for Correspondence: roba.bairakdar@mail.concordia.ca

1. INTRODUCTION

Wildfires burn around 423 million hectares of the global land surface annually, which represents approximately 3% of the global vegetation land (Giglio et al. 2018). Under very dry conditions, such as extreme heat combined with very low precipitations, a natural spark or human activity can ignite a wildfire. The effect of climate on wildfire risk varies depending on the location and time due to the spatial and temporal climate diversity, spatial variation of vegetation, fire management and prevention policies, and the interaction between these variables. The 2016 wildfires in Alberta (Fort McMurray) and the 2017-2021 wildfires in California generated billions of dollars in insured losses and alerted the public and insurers to the threats caused by wildfires. Knowledge of the spatial likelihood of burning and the potential financial impact of wildfires can provide a great tool for wildfire risk management.

Although fire activity in North America fluctuates annually, recent years witnessed an increased annual number and size of forest fires in some locations compared to the

past few decades. In fact, the 2017 and 2018 wildfires in British Columbia are the largest wildfires to be recorded in the area since the 1950s, with a total of 1.2 and 1.4 million hectares burned, respectively (Hanes et al. 2019). The Dixie fire in 2021 is the second-largest wildfire in the history of California, with around 0.4 million hectares burned (CAL-FIRE 2022).

Wildfire losses fluctuate considerably, depending on the location of events. Between 2000-2019, 26% of the world's wildfires and 69% of the economic losses due to wildfires occurred in the United States of America (USA), with the most damage in California (Yaghmaei 2020). In California, numerous insurers have stopped writing homeowners' insurance policies in fire-prone areas, resulting in a transfer of the insurance business to expensive specialty markets (Groom 2015). The Canadian Interagency Forest Fire Center (CIFFC) reports 6,317 wildfires across Canada in 2021, with more than 4.2 million hectares of burned land (CIFFC 2021). The International Disaster Database reports that North America suffered from \$68.8 billion in total damages (adjusted to inflation) due to wildfires over the period be-

-
- a Ph.D., ASA, was a Ph.D. candidate in the Department of Mathematics and Statistics at Concordia University, and Associate of the Society of Actuaries. Her research interests include catastrophe insurance risk modeling, extreme value theory, and univariate and multivariate risk measures. Her work experience includes valuation, experience analysis, identification of optimal reinsurance arrangements at AXA-Egypt, wildfire insurance risk modeling at Aviva, and statistical modeling at the research center of St. Justine Hospital.
- b Ph.D., FSA, FCIA, is Professor of actuarial science at Université du Québec à Montréal (UQAM), and Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries. His research interests include catastrophe modelling of climate extremes for (re)insurance purposes, flood insurance in Québec and Canada and financial risk-sharing for climate extremes. He has published papers on these topics and has research partnerships with the public and private sectors on the impacts of climate change on governments, municipalities, (re)insurance industry and pension plans.
- c Ph.D., ASA, ACIA, is Associate Professor in the Department of Mathematics and Statistics at Concordia University. Melina's research interests are related to modelling multivariate risks, related to property and casualty insurable risks. She is also interested in risk measures and uncertainty assessment. Several collaborations with the insurance industry and regulatory sector are in progress, related to natural catastrophes and extreme risks.

tween 2000–2019, with only 64% of those losses being insured (EM-DAT 2020).

Spatial wildfire risk assessment and fire spread models are growing fields with applications in fire-related decision making. Simulation-based fire spread algorithms were introduced to mitigate the damage caused by wildfires; see, for example, Van Wagtenonk (1996). Spatial wildfire models in the literature usually focus on a limited geographical study area; see, for example Finney et al. (2011) (continental USA), Atkinson et al. (2010) (Tasmania, Australia), Chuvieco et al. (2010) (Spain), Massada et al. (2009) (North-western Wisconsin, USA), and Van Wagtenonk (1996) (Sierra Nevada, USA), however, they are not created for actuarial purposes. Adequate estimation and prediction of future wildfire insured losses is challenging since it requires information on buildings. Moreover, constant developments in wildfire-prone areas may not be consistent with historical losses. As such, insurers are typically buying private complex wildfire insurance risk models, which may provide results that are difficult to interpret and replicate. Spatial modeling in the actuarial literature usually focuses on claims data; see, for example Shi and Shi (2017) for spatial frequency and severity modeling of auto-insurance claims, however, one may not always have sufficient claims data that covers the geographical region of interest. In the context of natural catastrophic events, such as wildfires, it is more suitable to follow a natural catastrophe risk modeling decomposition approach, i.e. identify the hazard, vulnerability and exposure; see Mitchell-Wallace et al. (2017) for details. There is wildfire risk at the intersection of those three components, such that a vulnerable property is subject to wildfire risk due to its location and characteristics.

This article provides a quantitative approach that considers the spatial variation in climate, land and demographic factors that drive wildfire occurrence in North America. We propose a wildfire risk model specifically tailored for USA and Canada, separately, splitting for each cause of ignition: man-made and natural. Our goal is to provide an open-source model, for pricing and underwriting P&C insurance policies. The model takes into account multiple types of wildfire risk features: continuous variables, categorical variables, and geographical coordinates. Several modeling techniques are used, such as GLMs and statistical learning algorithms that rely on decision trees. Combining numerous decision trees in an ensemble, such as a random forest, provides high predictive performance with an ability to discover interaction effects between the predictors. Tree-based machine learning models offer strong predictive capabilities, in addition to their interpretability and ease of explanation of the importance of the predictors, unlike neural network. They can also implicitly handle variable correlation, capture variable interactions and non-linear relationships between the predictors. Given the lack of publicly available wildfire claims data for each location in North America, we instead focus on modeling the annual burn probabilities by relying on publicly-available historical wildfire data. The goal of the article is to provide an interpretable model that predicts the annual burn probabilities for any location in North America and to

compare how random forest models perform compared to more commonly used actuarial modeling techniques, such as GLMs. The best fit model is used in a downscaling exercise where we predict the annual burn probabilities at a high resolution, which is applicable for insurance purposes, such as pricing home insurance policies with coverage against wildfires.

Wildfire risk in North America is of great significance and over the past decade, insurers have been suffering annually from billions of dollars in losses. This article proposes a transparent and simple, yet powerful model that relies on random forests. It can be possibly used by actuaries to price, reserve or manage the financial risk from wildfires. Unlike wildfire models that were designed for a specific region, the models presented in this article cover a wide geographic area. They have high predictive capabilities with strong ability to classify locations into high/low risk zones. The random forest models are characterised by their ability to capture non-linearity and interaction in the model inputs. Occurrence of wildfire can be devastating, thus it is of utmost importance to measure wildfire risk by calculating the burn probabilities of wildfires in high-risk areas with densely populated communities. With sufficient property characteristics data obtained at the underwriting stage, our model can be used to compute the loss costs necessary for premium calculation, accordingly, complementing vendor catastrophe-loss models.

The remainder of the article is organized as follows. Section 2 explains the data used in the model, such as the historical wildfires data, vegetation information and land use, meteorological and population data. It also classifies the population of North America depending on the wildfire risk. Section 3 provides details on the methods proposed to fit the data. It also summarizes the results of the fitted models and compares their predictive capabilities. Section 4 provides a detailed analysis of the effect of each input variable. An insurance application is illustrated in Section 5 to underwrite and price insurance policies. Finally, Section 6 concludes the article.

2. DATA

In this section, we discuss the data used to build our model. We explain the historical wildfires data and the wildfire risk factors such as climate, land cover, population census, and lightning frequency.

2.1. WILDFIRE OCCURRENCE AND INTENSITY

The area burned and the frequency of wildfires are relevant metrics to identify trends in wildfire risk and they are also of great importance to estimate insurance losses. Summaries of the historical wildfires in North America are provided in [Figure 1](#), where the data is obtained from (CFS 2019; Short 2017). Data for Canadian wildfires range from 1975 until 2019. 2016 witnessed 1.3 million hectares burned - the Fort McMurray wildfire being responsible for 0.6 million. The last five years of data (2015–2019) had 29,000 wildfires causing a total of 13 million hectares

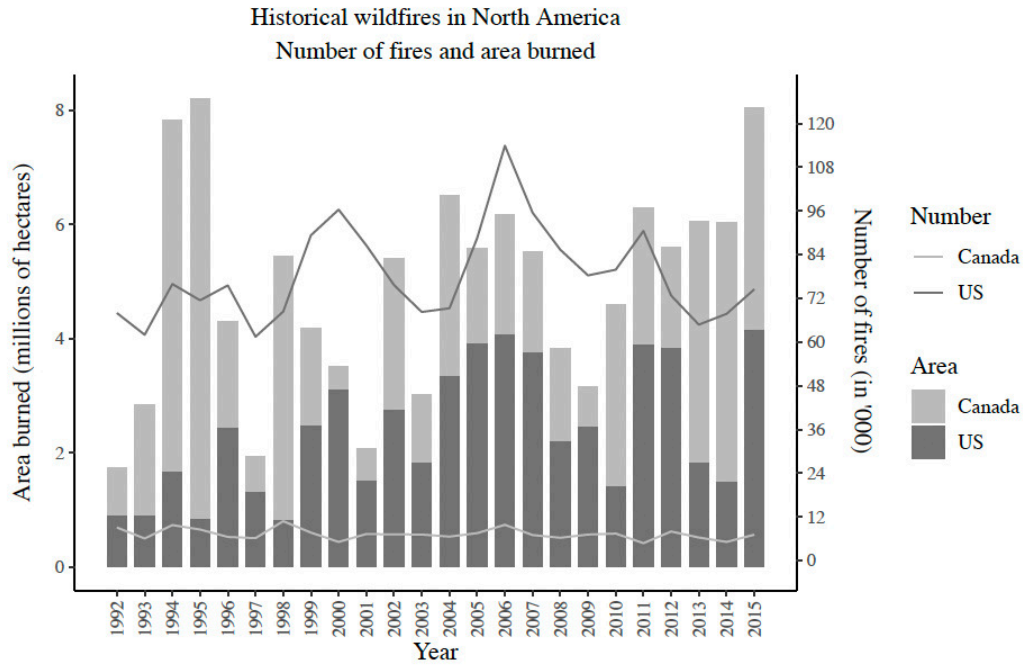


Figure 1. The annual number of wildfires (lines) and area burned (bars) in North America.

Each bar shows the split of the total area burned over the USA and Canada.

burned, with an average fire size of 450 hectares. Data for the USA wildfires range from 1992 until 2015. The last five years of data (2011–2015) had 370,000 wildfires resulting in a total of 15 million hectares burned, with an average fire size of 40 hectares. The increase in the area burned can be attributed to the lengthening of fire seasons due to the increase in spring and summer temperatures, as explained by Westerling et al. (2006). Historical wildfires in the USA compared to Canada contain a substantial number of wildfires that are small in size, however, Canadian wildfires data contain relatively larger fires. This may be explained by the differences in reporting of wildfires by the fire agencies.

Lightning strikes are the leading cause of large wildfires, which is typically due to prolonged dry conditions, high temperatures, and excessive dry fuel. However, in densely populated areas, man-made sparks can be responsible for a large number of ignitions due to human-related activities, such as arson, campfires, equipment use, fireworks, smoking, etc. [Figure 2](#) compares the annual number and size of wildfires of all sizes caused by lightning strikes and human ignitions, where we observe that natural wildfires are significantly less frequent compared to man-made wildfires, but they tend to burn more land and cause more damage.

Numerous small wildfires may not necessarily result in a significant increase in the area burned. Accordingly, it may be more appropriate to only consider large wildfires as they are more likely to cause material damage. The Canadian large wildfire database includes information on all wildfires larger than 200 hectares (Stocks et al. 2002). The large wildfires in Canada represent 4% of the count of all wildfires in Canada and 98.8% of the total area burned in the observed period. 84% of the Canadian large wildfires are ignited by lightning strikes and they contribute to 91% of the

total area burned by all large wildfires in Canada. By applying the same threshold of 200 hectares, less than 1% of the count of wildfires in the USA is considered large and they correspond to 88.8% of the total area burned. 44.6% of the large wildfires in the USA are ignited by natural reasons and they contribute to 67.8% of the total area burned.

[Figure 3](#) maps the historical annual burn probabilities for each cause per grid cell. The chosen resolution is $0.1^\circ \times 0.1^\circ$ where 1° is approximately 111 km at the Equator, which is the resolution of the meteorological data. The count of wildfire origins that occurred in each cell is calculated for the entire duration of each dataset, 45 years for Canada and 24 years for the USA. Each year is coded as binary because only a few grid cells have more than one wildfire in a given year. The annual burn probability is obtained by dividing by the number of years of data. Due to the different natures of the maps, we choose to split our North American wildfire model by country and by cause. Wildfire risk is heterogeneous in space, which means that observations are not identically distributed, i.e. in each grid cell, there might be a different distribution. In addition, one could have spatial dependence in between the random variables, i.e. if there is a claim in a grid cell, then it is highly likely there is a claim in the neighbouring grid cell.

2.2. PREDICTORS/DETERMINANTS OF WILDFIRES

[Table 1](#) summarizes the wildfire risk predictors explained in this section and used throughout this paper. All meteorological data is obtained from the Copernicus Climate Change Service Climate Data Store (CDS) ERA5 reanalysis data at a spatial resolution of $0.1^\circ \times 0.1^\circ$ (CDS 2019). Monthly-mean averages for the observed period are ob-

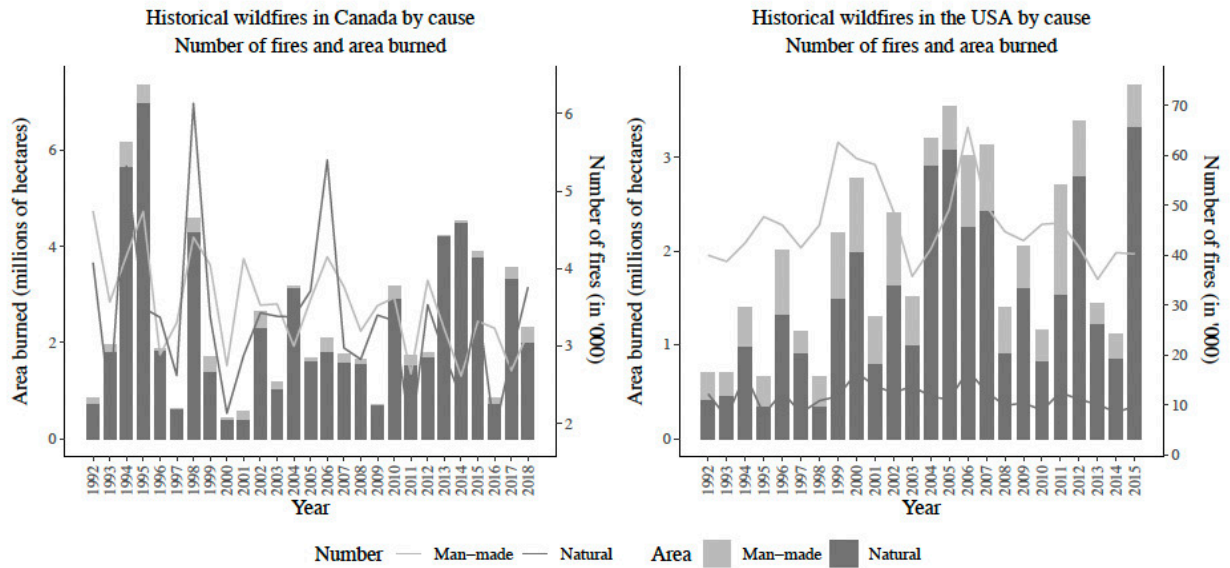


Figure 2. Annual number of wildfires (lines) and area burned (bars) by cause of wildfire in North America.

Each bar shows the split of the total area burned by the cause of the wildfire; i.e. man-made vs. natural wildfires.

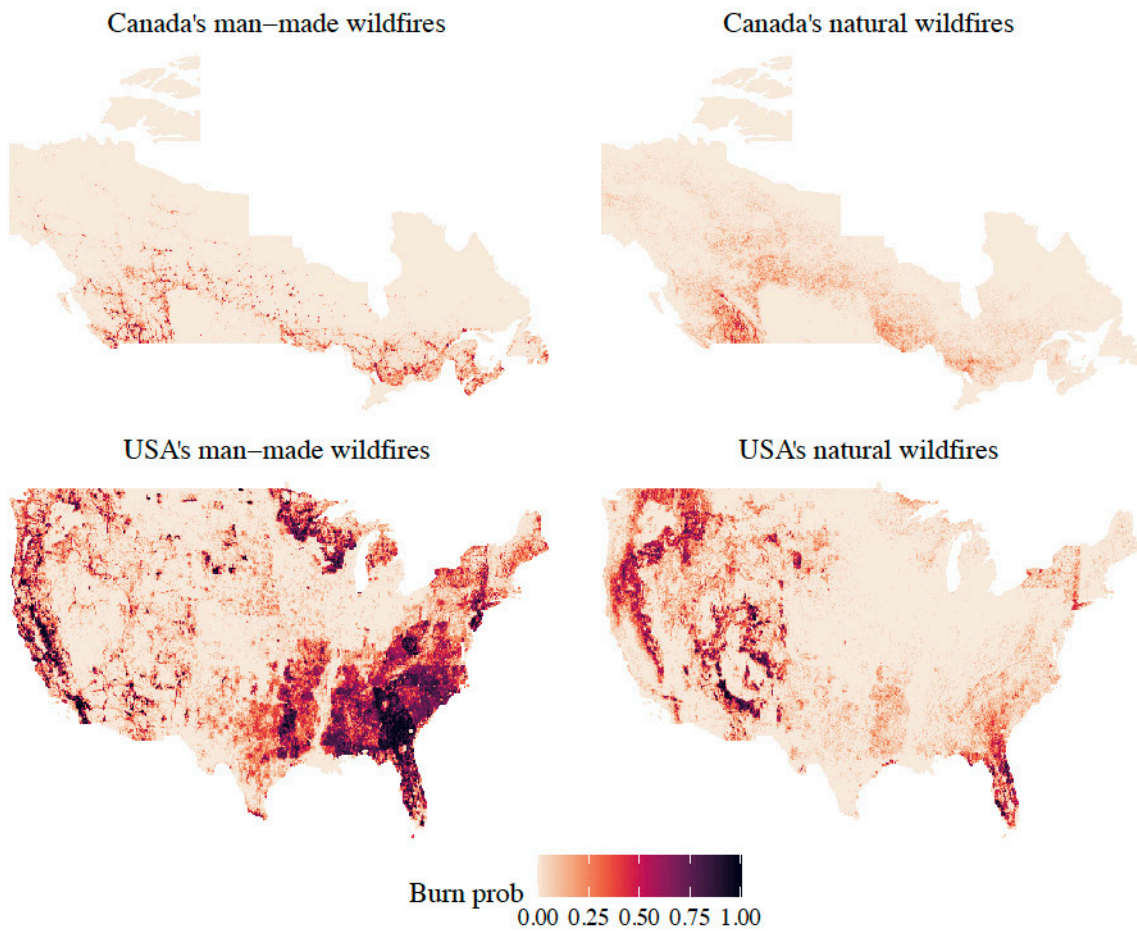


Figure 3. Observed annual burn probabilities per grid cell.

The maps are split by country and cause of wildfire.

Table 1. Definitions of the modeling parameters.

Variable	Definition
Longitude	The longitude at the center of the grid cell
Latitude	The latitude at the center of the grid cell
LandCover_1	Percentage of the grid cell land covered by temperate or sub-polar needleleaf forest
LandCover_2	Percentage of the grid cell land covered by sub-polar taiga needleleaf forest
LandCover_3	Percentage of the grid cell land covered by tropical or sub-tropical broadleaf evergreen forest
LandCover_4	Percentage of the grid cell land covered by tropical or sub-tropical broadleaf deciduous fores
LandCover_5	Percentage of the grid cell land covered by temperate or sub-polar broadleaf deciduous fores
LandCover_6	Percentage of the grid cell land covered by mixed forest
LandCover_7	Percentage of the grid cell land covered by tropical or sub-tropical shrubland
LandCover_8	Percentage of the grid cell land covered by temperate or sub-polar shrubland
LandCover_9	Percentage of the grid cell land covered by tropical or sub-tropical grassland
LandCover_10	Percentage of the grid cell land covered by temperate or sub-polar grassland
LandCover_11	Percentage of the grid cell land covered by sub-polar or polar shrubland-lichen-moss
LandCover_12	Percentage of the grid cell land covered by sub-polar or polar grassland-lichen-moss
LandCover_13	Percentage of the grid cell land covered by sub-polar or polar barren-lichen-moss
LandCover_14	Percentage of the grid cell land covered by wetland
LandCover_15	Percentage of the grid cell land covered by cropland
LandCover_16	Percentage of the grid cell land covered by barren lands
LandCover_17	Percentage of the grid cell land covered by urban
LandCover_18	Percentage of the grid cell land covered by water
LandCover_19	Percentage of the grid cell land covered by snow and ice
Elevation	Height above sea level
Population	Count of individuals who reside in the grid cell
LightningFrequency	The annual frequency of lightning strikes per grid cell
Temperature_FireSeason	The average daily temperature during the fire season
Temperature_NotFireSeason	The average daily temperature in days falling outside the fire season
TotalPrecipitation_FireSeason	The average daily total precipitation during the fire season
TotalPrecipitation_NotFireSeason	The average daily total precipitation in days falling outside the fire season
EastwardWind_FireSeason	The average speed of air moving towards the east during the fire season
NorthwardWind_FireSeason	The average speed of air moving towards the north during the fire season

tained for the following variables: (1) the air temperature at 2m above the surface of the Earth, (2) the horizontal speed of air moving towards the east and the north at a height of 10m above the surface of the Earth, and (3) the total precipitation from the accumulated liquid and frozen water, including rain and snow that falls to the surface of the Earth. The data is then averaged by fire season, where fire season is defined to be March to September for the USA wildfires, May to September for Canada's natural wildfires, and April to October for Canada's man-made wildfires. The fire seasons are obtained from the historical wildfire data by observing the monthly seasonality for each wildfire cause. All other covariate information are aggregated/re-scaled to $0.1^\circ \times 0.1^\circ$ to unify the resolution and are stored as raster images, which are pixel-based files that contain unique information per cell.

We use the 2015 North American Land Cover 30m dataset (NALCMS 2020), which explains the material features of the Earth's surface at a 30m spatial resolution

based on Landsat-7 imagery. To unify the resolution with the other covariates, the percentage of each land type is calculated for each $0.1^\circ \times 0.1^\circ$ grid cell, hence, converting the 19 types of land cover from a categorical variable to 19 continuous variables that range from 0 to 1. The world digital elevation model, created from a digital database of land and sea-floor elevations on a $1^\circ/12$ resolution, is used as a predictor that represents elevation (NOAA 2016). The global map of lightning frequency contains the number of strikes/km²/year on a $0.5^\circ \times 0.5^\circ$ grid. In North America, Florida has the highest annual lightning strike rate, with an average of 59 strikes/km²/year (Cecil, Buechler, and Blakeslee 2014).

Population count data is obtained from the fourth version of the Gridded Population of the World collection, which provides a smoothed distribution of the human population on a continuous global raster surface at a resolution of $1^\circ/120 \times 1^\circ/120 \approx 1 \text{ km} \times 1 \text{ km}$ (CIESIN 2017). The dataset is created from census data collected between 2005

and 2014 and then extrapolated to estimate the population count/density for the years 2000, 2005, 2010, 2015, and 2020. For this article, we use the raster for the population count in 2005 as a possible predictor and the population count in 2020 is used for predictions.

2.3. POPULATION AT RISK

To analyze the population exposed to wildfire risk, we classify the population in each state and province into low, medium or high risk based on the observed burn probabilities of natural or man-made wildfires. We define low risk to be an annual likelihood $< 10\%$, medium risk is between $10 - 40\%$ and high risk are for grid cells above 40% . [Table 2](#) summarizes the results. Overall, 8% of Canadian residents are at high risk of man-made wildfires, compared to 33% of US residents. While no Canadian residents are located in high risk zones for natural wildfires, 5% of the USA population resides in zones that are at high risk for natural wildfires.

3. WILDFIRE OCCURRENCE MODELS

This section explains the models used in this paper and then provides the results of the fitted models and compares their predictive capabilities. We are interested in modeling N , the random number of wildfires for each location in North America over a grid of $0.1^\circ \times 0.1^\circ$, over a time period t ; 45 years for Canada and 24 years for the USA. Due to the significant natural climate variability, we use as much available data over time as possible to cover most possible climate scenarios. The annual burn probability per grid cell is N/t . Following the discussion of determinants of wildfires in Section 2.2, the predictors of the models are: geographical location, percentage coverage over a grid cell by each land type, elevation, population count, lightning frequency, the mean temperature in fire season, the mean temperature other than fire season, the mean precipitation in fire season, the mean precipitation other than fire season, the mean eastward wind speed in fire season and the mean northward wind speed in fire season. See [Table 1](#) for the definitions of the predictors.

Wildfire occurrence models are conditional on meteorological, topographical and socioeconomic variables, aiming to find the relationship between those sets of covariates and wildfire occurrence, as such, any trend in wildfire occurrence is implicitly captured by the trend in the covariates.

3.1. METHODOLOGY

We compare GLMs (McCullagh 1984), regression decision trees (Breiman et al. 2001) and random forests (Breiman 2001). Even though interpreting GLMs is quite simpler than tree-based models, yet the latter can capture interactions easily and model non-linear relationships between predictors. Decision trees provide predictions that are easily interpreted, however they are known to have high variance, where small changes in the data can generate different

trees and hence different predictions for some observations. This limitation is taken care of by using random forests, which relies on numerous decision trees, thus minimizing the variance.

To fit the models and compare their predictive power, we first randomly partition the data into train (90%) and hold-out test (10%) datasets. We apply repeated 10-fold cross-validation (CV) with 5 repetitions to train the candidate models and tune the hyperparameters of the tree-based models. As a model validation technique, CV allows us to assess how the performance of our models can generalize to other independent datasets. In each iteration of the CV, we use 9 different folds out of the 10 folds to train our model, and we test on the remaining fold. We evaluate the performance of the tree-based models over an extensive grid search of possible values of the hyperparameters. The best set of hyperparameters are the ones that minimize the RMSE. See Bruce and Bruce (2017) for further details. The hold-out test dataset is later used to compare predictions with the actual observed values. The *Caret* (Kuhn 2021) R package is used to tune the hyperparameters, train the models and calculate variable importance.

In the next section, we present the details of the fitted models. First, we model the occurrences of wildfires of any size, split by country and cause of wildfire. Then, we consider large wildfires only and we model them by country.

3.2. MODELS FOR WILDFIRES OF ANY SIZE

The wildfires are split by ignition cause; man-made or natural. Accordingly, we fit four models:

1. Canada man-made wildfires,
2. Canada natural wildfires,
3. USA man-made wildfires, and
4. USA natural wildfires.

[Figure 4](#) displays the distribution of the response variable in each data set, representing generally balanced datasets. Canada man-made wildfires had a maximum of 44 wildfires in a grid cell over 45 years, i.e. almost a wildfire occurred in approximately the same location every year. However, Canada's natural wildfires had a maximum of 32 wildfires in a grid cell over 45 years. The USA man-made and natural wildfires each had a maximum of 24 wildfires over 24 years.

We apply the algorithms discussed in Section 3.1 to train the data and tune the hyperparameters of the tree-based models. To assess the predictive accuracy of the models, we calculate the CV root mean squared prediction error (RMSE) and the CV mean absolute prediction error (MAE). To facilitate the comparison between the models for Canada and the USA, standardized versions of the RMSE and MAE, that account for the number of years in each model, are calculated. Thus, both measures are presented in terms of the annual burn probability. For model comparison purposes, we also compare the CV Pearson correlation coefficient between the observed values and the predictions, and we compute the RMSE on the hold-out test dataset. Additionally, we calculate the prediction error measures for a baseline model, which assumes that the

Table 2. Wildfire risk classification by population. The states and provinces are ordered in descending order of their count of wildfires. The top 5 provinces in Canada and the top 10 states in the USA are shown.

Man-made					Natural				
Province	Number of wildfires ('000)	Low risk (%)	Medium risk (%)	High risk (%)	Province	Number of wildfires ('000)	Low risk (%)	Medium risk (%)	High risk (%)
Canada	179	75	17	8	Canada	153	97	3	0
British Columbia	42	49	25	26	British Columbia	51	79	19	2
Ontario	30	93	5	2	Ontario	28	99	1	0
Alberta	28	92	4	4	Alberta	25	98	2	0
Québec	27	56	34	10	Saskatchewan	11	98	2	0
Saskatchewan	13	90	3	7	Québec	11	100	0	0
USA	1105	43	25	33	USA	274	85	10	5
Georgia	149	6	14	80	Oregon	32	84	10	5
California	97	27	33	40	Arizona	29	90	6	4
Texas	93	37	49	14	California	27	93	6	1
North Carolina	80	2	9	89	Florida	22	41	34	25
Mississippi	60	6	17	77	Idaho	21	83	15	1
Alabama	55	0	23	77	Colorado	20	80	15	5
Florida	54	12	17	70	New Mexico	18	70	26	4
South Carolina	49	1	12	87	Montana	18	66	28	6
Minnesota	37	59	17	24	Utah	17	64	30	6
Tennessee	29	40	34	26	Nevada	11	78	19	3

Low risk: annual likelihood < 10%, medium risk: annual likelihood 10 - 40%, and high risk: annual likelihood > 40%.

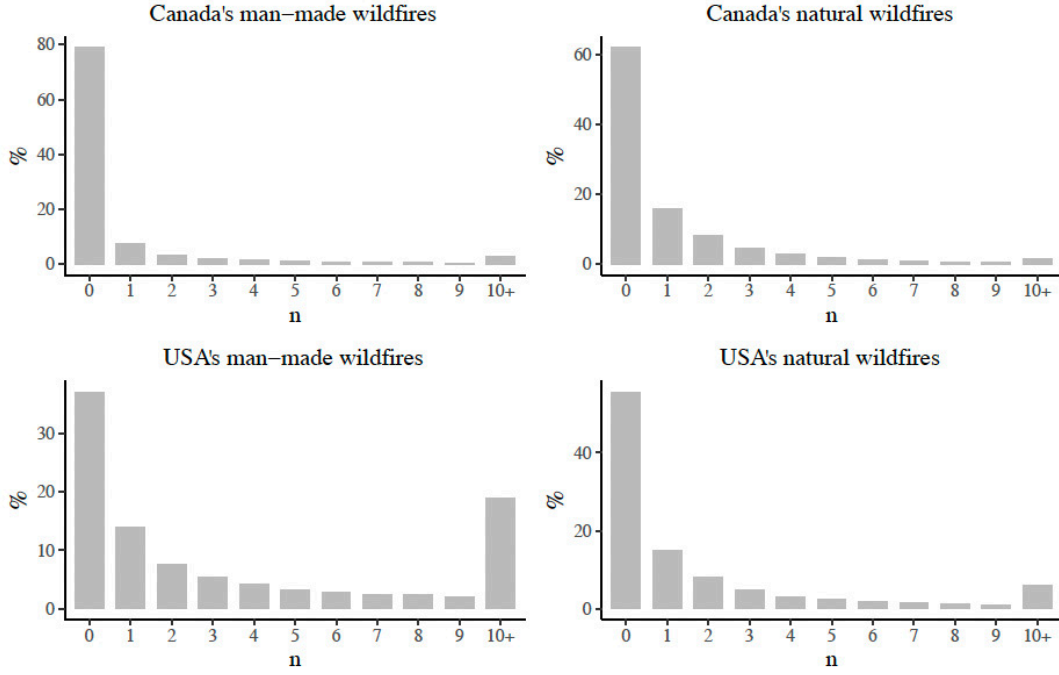


Figure 4. The proportion of $0.1^\circ \times 0.1^\circ$ grid cells with n wildfires of any size in 45 years in Canada and 24 years in the USA.

country-wide mean is applied to all grid cells. Table 3 summarizes the discussed measures. As expected, the random forest models outperform the other fitted models. We also observe that the random forest models show around half the prediction error of the baseline model.

The random forest models have the strongest predictive power, and hence they are chosen as the optimal models such that all the analysis thereafter is performed on them. As a comparison tool against historical wildfire records, Figure 5 provides a visual representation of the differences between the predictions of the random forest models and the observed events for each grid cell to highlight zones where the model is lacking predictive power. To facilitate the comparison between the four maps, the differences are computed in terms of the annual burn probability.

Figure 6 displays the variable importance for each random forest model, which represents the percentage increase in the mean square error of the random forest model when the data for that variable are randomly permuted (Breiman et al. 2001). The analysis shows that for man-made wildfires in Canada and the USA, the most important predictors are population, urban land, water and northward wind speed. For natural wildfires, there are no common important predictors between the countries.

Figure 7 compares the predicted annual burn probabilities to the observed burn probabilities in the hold-out test dataset, where we notice a strong linear relationship close to the 45° line, except for slight under prediction of high-risks. This is generally acceptable due to the difficulty of any model in predicting outliers/extremes. For each observation, prediction intervals are plotted, which represent 95% of the individual predictions from the decision trees

composing each random forest model. In fact, around 90% of the observed burn probabilities for the USA wildfires and 95% of the observed burn probabilities for Canadian wildfires fall within the 45th and 55th quantiles of the individual decision tree predictions.

3.3. MODELS FOR LARGE WILDFIRES

Even though large wildfires are small in number, they contribute significantly to the area burned and economic damage. Figure 8 displays the distribution of the response variable in each model. There was a maximum of 10 wildfires in a grid cell over 45 years in Canada, compared to 15 large wildfires in the USA over 24 years. We fit two models:

1. Canada large wildfires, and
2. USA large wildfires.

See Table 1 for the predictors of the models. We apply the methods discussed in Section 3.1 to train the data and tune the hyperparameters. Table 4 provides the CV RMSE, MAE and Pearson correlation coefficient of the fitted models, in addition to the RMSE computed on the hold-out test dataset. Both RMSE and MAE are presented in terms of the annual burn probability. Similar to the models in Section 3.2, the random forest models outperform the GLM and decision tree models. It is observed that the Pearson correlation coefficient of both models is low in comparison with the models for wildfires of any size, discussed in Section 3.2, which is attributed to the difficulty in predicting the large observation values.

As shown in Figure 8, where we have a large number of grids with no observed large wildfires, it is easier for

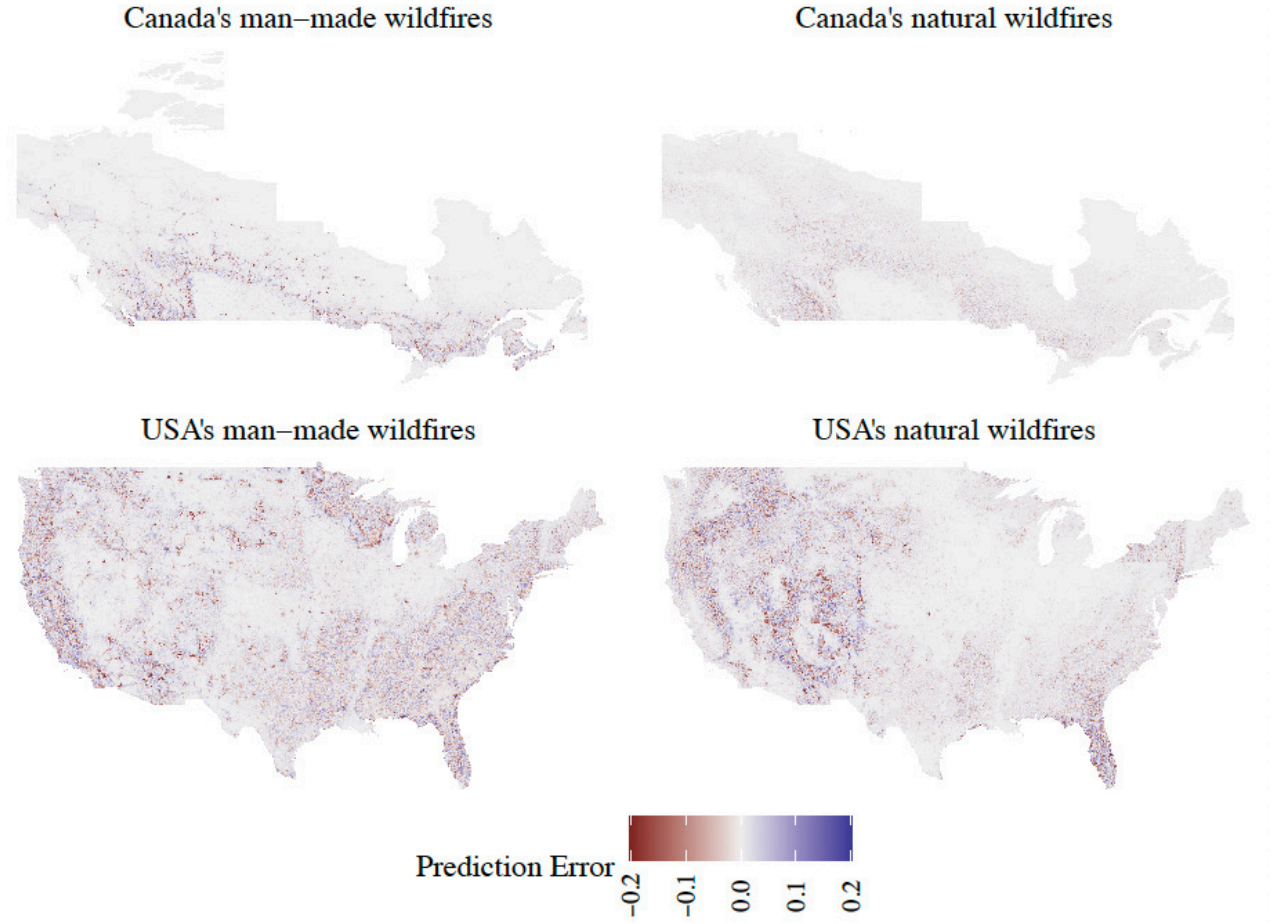


Figure 5. Difference between the predicted and actual annual burn probabilities of wildfires of any size.

the model to predict the zeros. Accordingly, we consider hurdle models, first introduced by Cragg 1971, which are pure mixtures of zero and non-zero outcomes. A logistic GLM is responsible for the binary component of whether the outcome is zero or positive. If the outcome is positive, the conditional distribution of the non-zeros is modeled by a zero-truncated count distribution. We apply the same methodology by using two-stages random forest model: the first stage is a classification model that categorizes the observations into “no large fire” vs “large fire” occurrence, while the second stage is a regression model to predict the number of occurrences should at least one large wildfire happens. In the first stage, we apply larger weights on the observations with “large fire” so that they have higher chances of selection in the bootstraps samples for the individual decision trees.

We assess the predictive strength of each model component separately: the hurdle component and the count component. For the hurdle component, we calculate the CV area under the receiver operating characteristic curve (AUC) and accuracy, i.e. the ratio of the correctly predicted observations to the total number of observations. While for the count component, we calculate the CV RMSE and Pearson correlation coefficient of the observed count of large wildfires and predicted count of large wildfires. The RMSE is presented in terms of the annual burn probability.

Finally, we compute the same measures on the hold-out dataset. [Table 5](#) summarizes the discussed measures. For the hurdle components, all attempted models perform similarly. By observing the AUC, the random forest models are slightly better, however by comparing the accuracy of the models, the GLM hurdle models seem to perform better. For the count components, the random forest models outperform the GLM hurdle models that fail to predict values in the tail of the distribution.

Overall, the two-stages random forest models have the strongest predictive power, and hence they are chosen as the optimal models for the occurrence of large wildfires in Canada and the USA. All the analysis thereafter for large wildfires is performed on the two-stages random forest models. [Figure 9](#) illustrates the difference between the predictions of the random forest models for large wildfires and the observed events for each grid cell, where we observe minimal errors, comparable to the small RMSE values shown in [Table 5](#). The values are computed in terms of the annual burn probability.

[Figure 10](#) displays the variable importance for each model in the two-stages random forest models to help identify the most influential variables. The comparison between the variables shows that climate variables are the main drivers behind wildfires becoming large in size.

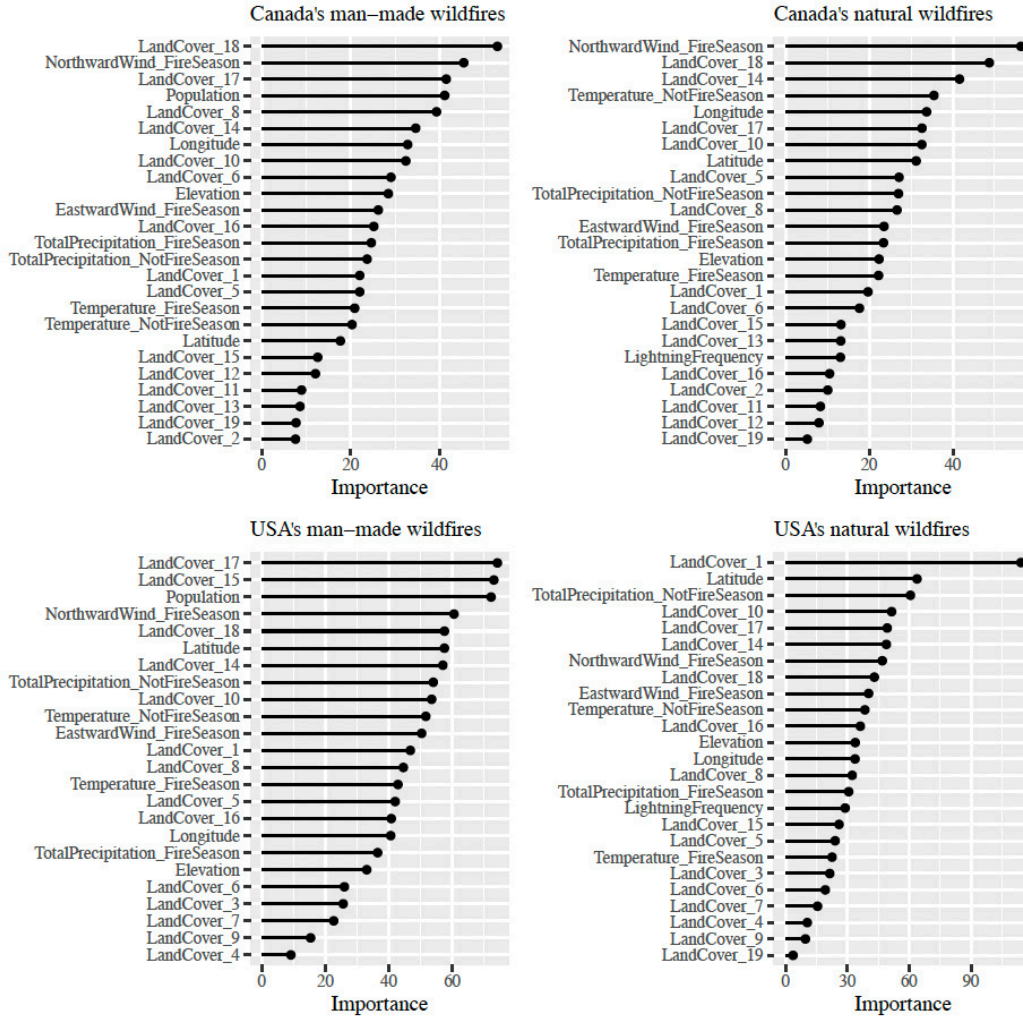


Figure 6. Variable importance (percentage mean decrease in accuracy) of the random forest models for wildfires of any size for each country and cause.

4. SENSITIVITY ANALYSIS

The wildfire random forest models have a large number of input variables, with their importance ranked in Figures 6 and 10. Note that variable importance does not quantify the effect of changes in the predictors on wildfire risk. In this section, we compute that effect by performing sensitivity tests on some of the predictors, while keeping all other model components unchanged. Table 1 contains a list of the performed sensitivities. For each sensitivity, a shock is applied to one (or more) predictor(s), and predictions of the random forest models that are trained in Section 3 are calculated. The analyses are evaluated over British Columbia, Alberta and California, which are chosen because of their high level of wildfire risk.

Tables 7 and 8 provide some summary statistics of the predicted annual likelihood of a wildfire of any size and of large wildfires for each sensitivity, to be compared with the base scenario, where all predictors remain unchanged. The calculated statistics are the mean, standard deviation and percentiles of the burn probabilities over all grid cells in each state and province. More percentiles in the tail are

provided for the sensitivities of large wildfires to reflect the behavior in the table. The burn probability of wildfires of any size is the sum of two components; the predicted likelihood of a wildfire due to natural causes multiplied by the historical probability of lightning-caused wildfires and the predicted likelihood of a fire due to man-made causes multiplied by the historical probability of man-made wildfires for each grid cell. The predicted annual likelihood of a wildfire of any size for each cause is obtained from the random forest models presented in Section 3.2. The results of the sensitivities are inline with the expectations; an increase in temperature, a decrease in total precipitation, an increase in lightning activity and human population are all factors that will increase the annual burn probabilities in the study regions. A change in the mean temperature has a stronger effect on the burn probabilities in British Columbia and Alberta compared to California. Even though a change in urban population affects the annual burn probabilities of wildfires of any size, it has a negligible effect on large wildfires because they are more commonly ignited from natural sparks.

Table 3. Cross-validation prediction errors and Pearson correlation coefficient and hold-out RMSE of the attempted models for fitting wildfires of any size for each country and cause. RMSE and MAE are presented in terms of the annual burn probability.

	Fitted model	Canada		USA	
		Man-made	Natural	Man-made	Natural
CV RMSE	Baseline	0.0706	0.0499	0.2586	0.1525
	GLM - Poisson	0.0617	0.0377	0.2032	0.1178
	GLM - Negative Binomial	0.0706	0.0399	0.2438	0.1264
	Regression Decision Tree	0.0580	0.0435	0.2179	0.1244
	Random Forest	0.0426	0.0267	0.1144	0.0722
CV MAE	Baseline	0.0343	0.0305	0.2036	0.0996
	GLM - Poisson	0.0234	0.0199	0.1296	0.0681
	GLM - Negative Binomial	0.0264	0.0201	0.1488	0.0702
	Regression Decision Tree	0.0244	0.0252	0.1565	0.0755
	Random Forest	0.0156	0.0148	0.0745	0.0407
CV Correlation	GLM - Poisson	0.4985	0.6582	0.6212	0.6372
	GLM - Negative Binomial	0.4555	0.6276	0.5219	0.6089
	Regression Decision Tree	0.5689	0.4873	0.5387	0.5788
	Random Forest	0.7988	0.8445	0.8980	0.8819
Hold-out RMSE	Baseline	0.0703	0.0501	0.2604	0.1525
	GLM - Poisson	0.0597	0.0371	0.2082	0.1191
	GLM - Negative Binomial	0.0721	0.0391	0.2524	0.1279
	Regression Decision Tree	0.0598	0.0457	0.2223	0.1290
	Random Forest	0.0416	0.0261	0.1154	0.0697

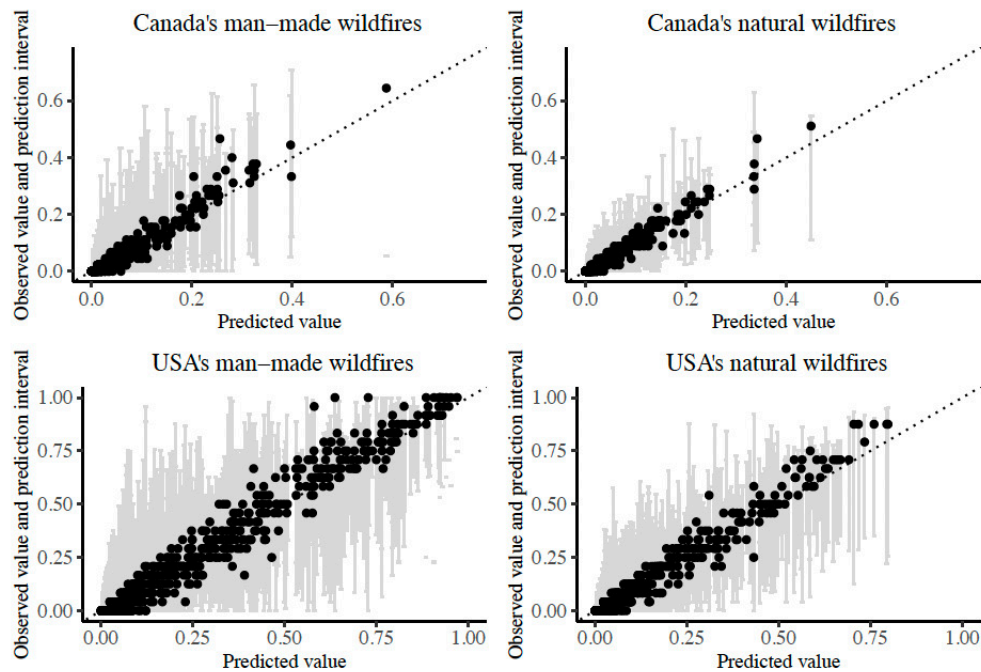


Figure 7. Predicted vs observed annual burn probabilities of wildfires of any size, computed on the hold-out test datasets.

The vertical bars are the prediction intervals.

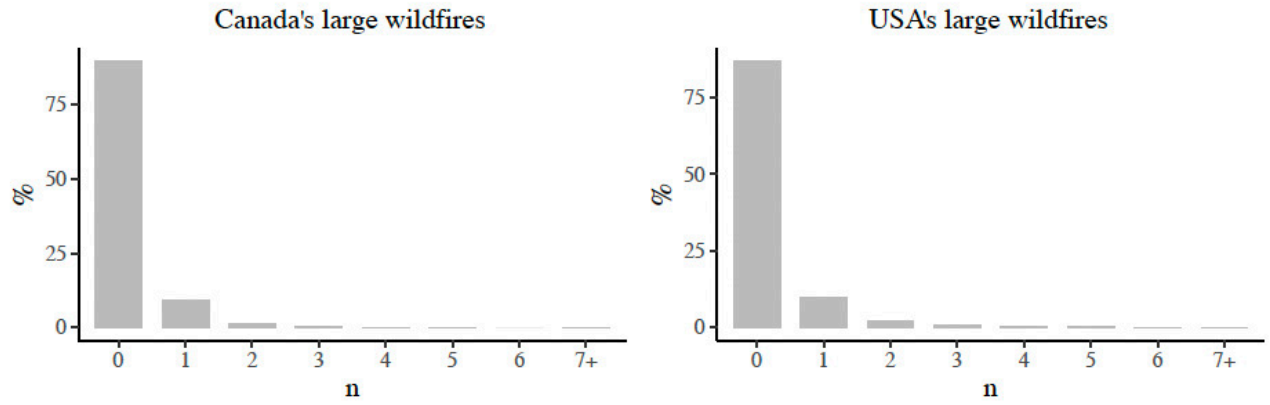


Figure 8. The proportion of $0.1^\circ \times 0.1^\circ$ grid cells with n large wildfires in 45 years in Canada and 24 years in the USA.

Table 4. Cross-validation prediction errors and Pearson correlation coefficient and hold-out RMSE of the models for fitting large wildfires for each country. RMSE and MAE are presented in terms of the annual burn probability.

	Fitted model	Canada	USA
CV RMSE	Baseline	0.0086	0.0244
	GLM - Poisson	0.0082	0.0232
	GLM - Negative Binomial	0.0082	0.0233
	Regression Decision Tree	0.0084	0.0235
	Random Forest	0.0079	0.0204
CV MAE	Baseline	0.0048	0.0136
	GLM - Poisson	0.0044	0.0122
	GLM - Negative Binomial	0.0044	0.0122
	Regression Decision Tree	0.0046	0.0125
	Random Forest	0.0042	0.0102
CV Correlation	GLM - Poisson	0.2714	0.2972
	GLM - Negative Binomial	0.2694	0.2956
	Regression Decision Tree	0.2142	0.2657
	Random Forest	0.3760	0.5471
Hold-out RMSE	Baseline	0.0082	0.0241
	GLM - Poisson	0.0079	0.0229
	GLM - Negative Binomial	0.0079	0.0229
	Regression Decision Tree	0.0080	0.0234
	Random Forest	0.0076	0.0198

We perform additional sensitivities to compare the distribution of the predicted annual burn probabilities of wildfires of any size over multiple scenarios of change in the mean temperature: $\pm 1^\circ$ and $\pm 2^\circ$. Figure 11 compares the densities of the sensitivities performed on the temperature predictor in the three chosen high-risk states and provinces. This figure complements the results from Table 7 by providing the full distribution across grid cells. One can see that across all study areas, an increase in temperature reduces the number of grid cells that have small predicted annual burn probabilities and increases the chances of medium and high wildfire risk, and vice versa. Addition-

ally, higher volatility can be observed in Alberta and California, compared to British Columbia.

5. INSURANCE APPLICATION

In this section, we aim to illustrate how we can use the models to price and underwrite wildfire risk in an insurance portfolio. The goal is to estimate the expected losses due to wildfire risk, based on the geographical location of the homeowner. A typical catastrophe model is composed of hazard modeling, exposure collection, vulnerability assessment and expected losses calculation; see Mitchell-Wallace

Table 5. Cross-validation and hold-out AUC, Accuracy, RMSE and Pearson correlation coefficient of the attempted hurdle models for fitting large wildfires for each country. RMSE is presented in terms of the annual burn probability.

		Fitted model	Canada	USA
Hurdle Component	CV AUC	GLM - Poisson Hurdle	0.7632	0.7458
		GLM - Negative Binomial Hurdle	0.7631	0.7459
		Random Forest	0.8213	0.8525
	CV Accuracy	GLM - Poisson Hurdle	0.8949	0.8685
		GLM - Negative Binomial Hurdle	0.8949	0.8686
		Random Forest	0.8757	0.8640
	Hold-out AUC	GLM - Poisson Hurdle	0.7494	0.7469
		GLM - Negative Binomial Hurdle	0.7494	0.7469
		Random Forest	0.8170	0.8539
	Hold-out Accuracy	GLM - Poisson Hurdle	0.8959	0.8685
		GLM - Negative Binomial Hurdle	0.8960	0.8686
		Random Forest	0.8749	0.8628
Count Component	CV RMSE	GLM - Poisson Hurdle	0.0233	0.0592
		GLM - Negative Binomial Hurdle	0.0234	0.0593
		Random Forest	0.0098	0.0336
	CV Correlation	GLM - Poisson Hurdle	0.1323	0.2297
		GLM - Negative Binomial Hurdle	0.1317	0.2249
		Random Forest	0.2414	0.4659
	Hold-out RMSE	GLM - Poisson Hurdle	0.0224	0.0581
		GLM - Negative Binomial Hurdle	0.0224	0.0582
		Random Forest	0.0052	0.0197
	Hold-out Correlation	GLM - Poisson Hurdle	0.1264	0.2310
		GLM - Negative Binomial Hurdle	0.1257	0.2264
		Random Forest	0.8809	0.8977

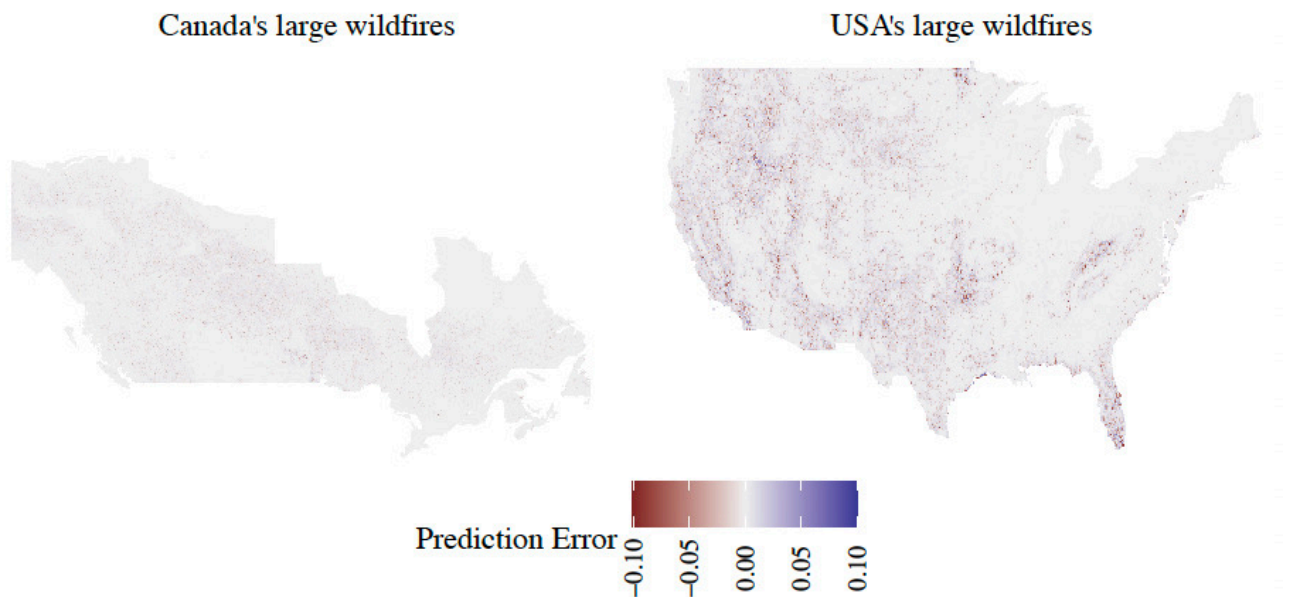


Figure 9. Difference between predicted and actual burn probabilities of large wildfires.

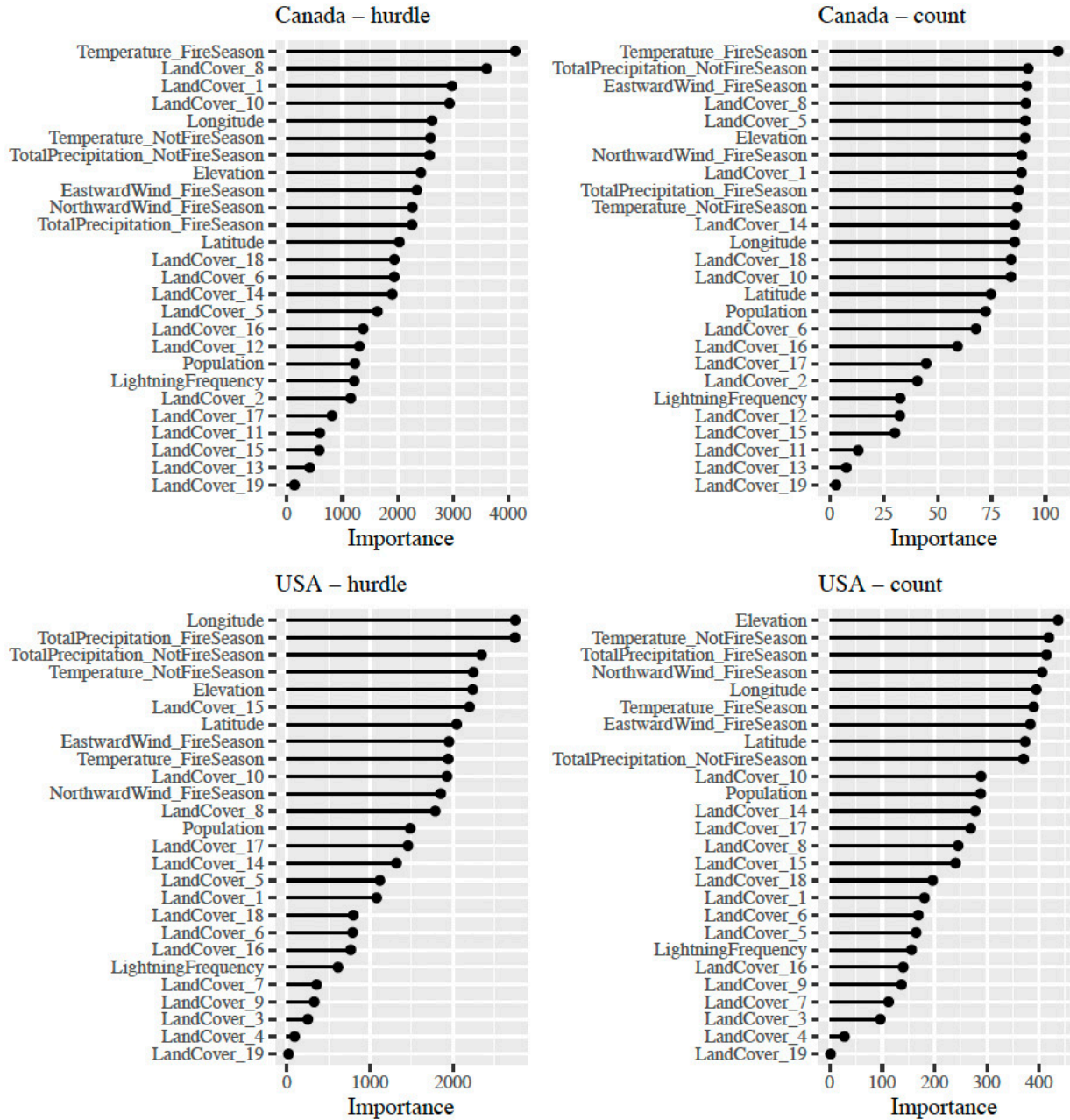


Figure 10. Variable importance (percentage mean decrease in accuracy) of the hurdle and count components of the random forest models for large wildfires for each country.

et al. (2017) for details. In this article, the hazard element is represented by the rate of wildfire occurrence, which varies across our chosen geographic region. This is reflected in the models built in Section 3 Collection of exposure requires the valuation of properties and/or infrastructures at risk. Vulnerability assessment helps in quantifying the relationship between the hazard and the damage by means of a metric, such as damage ratios. Finally, the annual pure premium is calculated in terms of the expected losses in each geographical location.

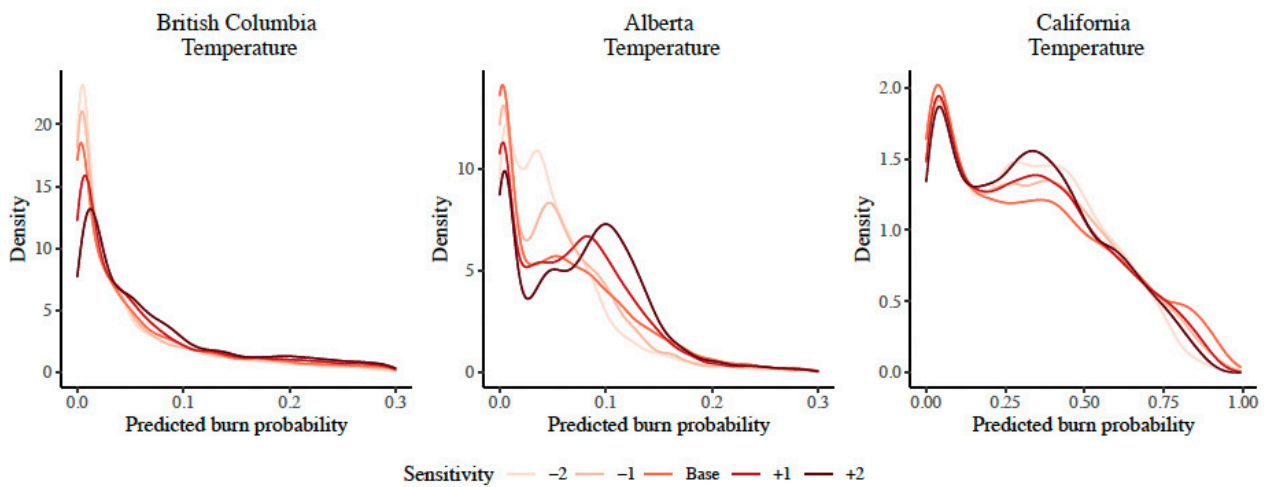
5.1. HAZARD

A grid cell of size $0.1^\circ \times 0.1^\circ$ may be too large for some insurance applications, such as pricing exercises, due to the need to understand the risk at a household level. As such, we perform a downscaling exercise where we predict the burn probabilities of a wildfire in a year at a much higher resolution. The chosen resolution is that of the Gridded Population of the World dataset, i.e. $1^\circ/120 \times 1^\circ/120 \approx 1 \text{ km} \times 1 \text{ km}$ grid cells, which is acceptable for insurers' purposes to accurately reflect the location of the insureds. A new high-resolution dataset is created by changing the resolution of all predictors, defined in [Table 1](#). For climate,

Table 6. Definitions of the sensitivities.

Sensitivity	Definition
Temp Inc	Increase the mean temperature by 1°.
Temp Dec	Decrease the mean temperature by 1°.
Precip Inc	Increase the mean total precipitation by 10%.
Precip Dec	Decrease the mean total precipitation by 10%.
Lightn Inc*	Increase the annual frequency of lightning strikes per km ² by 1 strike.
Lightn Dec	Decrease the annual frequency of lightning strikes per km ² by 1 strike.
Pop Inc	Increase population count and percentage of urban land by 10%.
Pop Dec	Decrease population count and percentage of urban land by 10%.

* An increase in the annual frequency of lightning strikes per km² by 1 strike represents an average increase of 103%, 40% and 84% in British Columbia, Alberta and California, respectively.

**Figure 11. Probability density function of the predicted annual burn probabilities of wildfires of any size for the sensitivity tests for the temperature predictors in British Columbia, Alberta and California.**

elevation and lightning strikes predictors, all small grid cells of size $1^\circ/120 \times 1^\circ/120$ within a large grid cell of size $0.1^\circ \times 0.1^\circ$ are assigned the same value. While for land cover predictors, the percentage of each land type is computed for each $1^\circ/120 \times 1^\circ/120$ grid cell. The random forest models in Sections 3.2 and 3.3 are used to predict the annual burn probabilities of wildfires of any size and of large wildfires per grid cell, by using the new high-resolution dataset.

The left panel in Figure 12 shows the burn probabilities of wildfires for each location in British Columbia by using the original dataset of resolution $0.1^\circ \times 0.1^\circ$ and by using the high-resolution dataset of size $1^\circ/120 \times 1^\circ/120$. Smaller grid cells indicate less number of predicted wildfires, and hence smaller burn probabilities. The annual burn probabilities in the small grid cells are scaled such that the overall likelihood of a wildfire is equivalent to that of the lower resolution model, while preserving the weights of the risk of the small grid cells. The right panel in Figure 12 shows a zoomed-in map on Lytton Creek and its surrounding region. As observed from the predictions of the original dataset and the downscaling exercise, the region has elevated burn probabilities compared to its surroundings and

to other areas in British Columbia. This coincides with Lytton Creek's large wildfire that occurred in 2021 in that area, which caused two fatalities and destroyed around 90% of Lytton village. The models were built on wildfires data up to the year 2019 in Canada, thus a 2021 wildfire is considered an out-of-sample prediction, hence confirming the appropriateness of the predictions of the models.

5.2. VULNERABILITY AND EXPOSURE

The average claim cost of a wildfire is highly volatile, depending on the location of the wildfire, the construction material of the buildings affected, their location, the fire mitigation measures and the availability of firefighters. Table 9 summarizes the average claim cost of the most recent large wildfires that occurred in three high-risk regions, British Columbia, Alberta and California. Thus, relying on such limited values to generalize over the whole region can be problematic. The unit of currency throughout is the Canadian dollar for Canadian wildfires and the US dollar for USA wildfires.

Vulnerability assessment for wildfire risk usually requires the help of civil engineers to analyze the character-

Table 7. Summary statistics of the predicted annual burn probabilities of wildfires of any size for the sensitivity tests.

State / Province	Statistic	Base	Temp Inc	Temp Dec	Precip Inc	Precip Dec	Lightn Inc	Lightn Dec	Pop Inc	Pop Dec
British Columbia	5 percentile	0.0004	0.0014	0.0007	0.0004	0.0006	0.0025	0.0004	0.0004	0.0004
	25 percentile	0.0037	0.0101	0.0050	0.0036	0.0055	0.0126	0.0036	0.0039	0.0037
	Mean	0.0696	0.0783	0.0612	0.0691	0.0709	0.0757	0.0651	0.0703	0.0687
	75 percentile	0.0907	0.1033	0.0795	0.0890	0.0926	0.0979	0.0866	0.0915	0.0900
	95 percentile	0.2840	0.2929	0.2490	0.2791	0.2847	0.2849	0.2473	0.2843	0.2805
	Standard deviation	0.0981	0.0998	0.0864	0.0952	0.0974	0.0956	0.0892	0.0983	0.0962
Alberta	5 percentile	0.0000	0.0003	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000
	25 percentile	0.0049	0.0161	0.0042	0.0091	0.0120	0.0051	0.0041	0.0053	0.0048
	Mean	0.0653	0.0709	0.0595	0.0654	0.0699	0.0679	0.0641	0.0659	0.0644
	75 percentile	0.0981	0.1038	0.0854	0.0989	0.1055	0.0998	0.0914	0.0991	0.0974
	95 percentile	0.1955	0.1978	0.1650	0.2008	0.1998	0.1976	0.1874	0.1960	0.1913
	Standard deviation	0.0726	0.0655	0.0620	0.0676	0.0676	0.0723	0.0698	0.0724	0.0705
California	5 percentile	0.0070	0.0107	0.0065	0.0068	0.0080	0.0081	0.0067	0.0072	0.0069
	25 percentile	0.0949	0.1135	0.0908	0.0940	0.1000	0.0990	0.0933	0.1009	0.0948
	Mean	0.3312	0.3364	0.3270	0.3358	0.3389	0.3334	0.3250	0.3333	0.3280
	75 percentile	0.5206	0.5345	0.4991	0.5157	0.5274	0.5298	0.5013	0.5248	0.5125
	95 percentile	0.8094	0.8113	0.7534	0.8003	0.8139	0.8099	0.8088	0.8091	0.8032
	Standard deviation	0.2561	0.2373	0.2365	0.2482	0.2465	0.2551	0.2530	0.2544	0.2531

Table 8. Summary statistics of the predicted annual burn probabilities of large wildfires for the sensitivity tests.

State / Province	Statistic	Base	Temp Inc	Temp Dec	Precip Inc	Precip Dec	Lightn Inc	Lightn Dec	Pop Inc	Pop Dec
British Columbia	Mean	0.0022	0.0022	0.0020	0.0022	0.0022	0.0023	0.0021	0.0022	0.0022
	90 percentile	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	95 percentile	0.0236	0.0244	0.0230	0.0233	0.0240	0.0243	0.0233	0.0241	0.0229
	97.5 percentile	0.0248	0.0255	0.0241	0.0240	0.0257	0.0253	0.0243	0.0254	0.0242
	99 percentile	0.0297	0.0311	0.0288	0.0294	0.0305	0.0308	0.0292	0.0310	0.0290
	Standard deviation	0.0072	0.0073	0.0069	0.0072	0.0072	0.0073	0.0071	0.0072	0.0072
Alberta	Mean	0.0027	0.0027	0.0027	0.0026	0.0027	0.0027	0.0027	0.0027	0.0027
	75 percentile	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	90 percentile	0.0230	0.0234	0.0227	0.0226	0.0234	0.0233	0.0228	0.0231	0.0229
	95 percentile	0.0244	0.0253	0.0240	0.0239	0.0253	0.0247	0.0241	0.0246	0.0243
	97.5 percentile	0.0257	0.0264	0.0251	0.0253	0.0265	0.0260	0.0252	0.0258	0.0254
	99 percentile	0.0338	0.0357	0.0315	0.0313	0.0345	0.0342	0.0313	0.0343	0.0331
	Standard deviation	0.0081	0.0081	0.0081	0.0082	0.0081	0.0081	0.0080	0.0081	0.0081
California	Mean	0.0172	0.0181	0.0169	0.0170	0.0173	0.0174	0.0171	0.0173	0.0170
	75 percentile	0.0471	0.0495	0.0465	0.0468	0.0485	0.0482	0.0467	0.0473	0.0468
	90 percentile	0.0603	0.0636	0.0586	0.0591	0.0620	0.0619	0.0586	0.0609	0.0591
	95 percentile	0.0734	0.0742	0.0728	0.0723	0.0739	0.0743	0.0727	0.0741	0.0723
	97.5 percentile	0.0865	0.0890	0.0844	0.0858	0.0883	0.0877	0.0842	0.0887	0.0846
	99 percentile	0.1067	0.1120	0.1048	0.1034	0.1106	0.1104	0.1029	0.1118	0.1042
	Standard deviation	0.0300	0.0301	0.0300	0.0301	0.0300	0.0300	0.0301	0.0300	0.0299

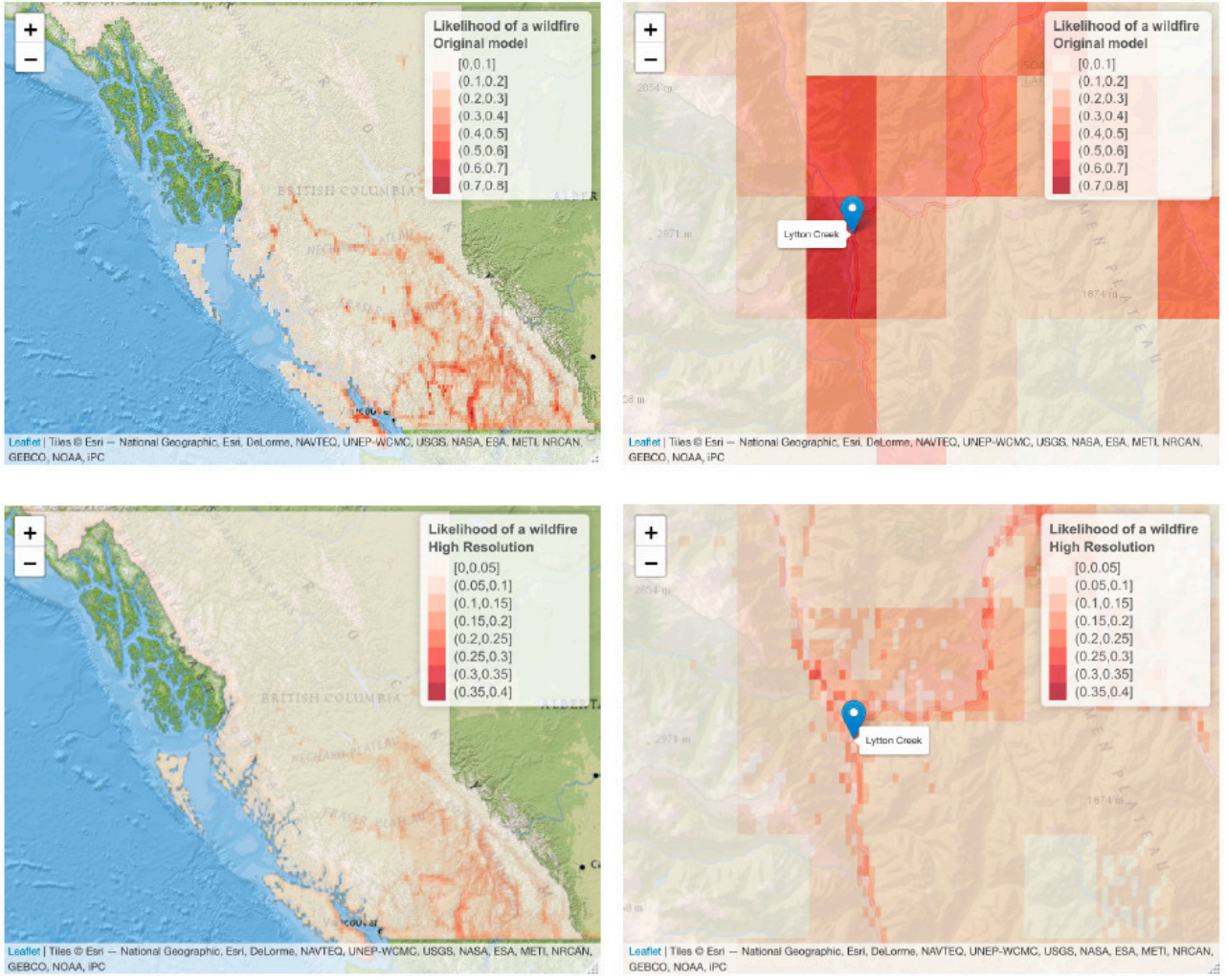


Figure 12. Predicted annual burn probabilities of wildfires of any size in British Columbia for the original (top) and high-resolution datasets (bottom), presented for the full province (left) and Lytton Creek (right).

istics of the properties. Those details are not publicly available, however, such information is typically provided by the customers to their insurers and is used in the calculation of the insured exposure. In this article, we rely on insured exposure data that is provided by CatIQ, Canada's Loss And Exposure Indices Provider (CatIQ 2021). They release annual updates of the Canadian insurance industry exposure database, which is developed from data from the Canadian P&C insurance companies. The 2020 year-end estimates of the personal properties sums insured and number of insured risks against fires is available for each Canada Post Forward Sortation Area (FSA). Depending on the population distribution in the FSA, the exposure value and the number of insured households in an FSA is distributed over each grid cell of size $1^\circ/120 \times 1^\circ/120$. Accordingly, the average exposure value per household is calculated for each grid cell.

5.3. INSURANCE PREMIUMS

Pure premium computation requires hazard, vulnerability and exposure calculations to be performed at the home-

owner's level. This entails detailed information on the location of the household, to determine its subjection to wildfire risk, in addition to property value and characteristics to evaluate its vulnerability if a wildfire occurs. In this section, we illustrate how the model can be used in such process.

We rely on the downscaled annual burn probabilities, provided in Section 5.1, to be the measure of hazard for each geographical location. With no access to vulnerability, we assume complete destruction of buildings and infrastructure where a large wildfire occurs, which will likely overestimate premiums. We choose to use the burn probabilities of large wildfire because insurance claims are more likely to occur from wildfires that are large in size, which cause complete destruction of the area. Hence, by using the CatIQ industry exposure database, the annual pure premium for wildfire insurance can be computed as the product of the annual burn probabilities of large wildfire by the cost of household replacement.

As an illustration of the technique, we analyze the distribution of pure premiums over British Columbia. The predicted insurance pure premium to cover wildfire risk in British Columbia is 0 for around 89% of the high resolution

Table 9. Historical average claim cost for some memorable large wildfires (CatIQ 2021; IBC 2021; CALFIRE 2021).

State/Province	Wildfire	Year	Average Claim Cost
British Columbia	White Rock Lake	2021	\$55k
	Lytton Creek	2021	\$253k
	Williams Lake	2017	\$15k
	Elephant Hill	2017	\$12k
Alberta	Fort McMurray	2016	\$64k
	Slave Lake	2011	\$110k
California	Multiple wildfires	2020	\$826k
		2019	\$574k
		2018	\$178k
		2017	\$1,116k
		2016	\$116k
		2015	\$958k
		2014	\$30k
		2013	\$60k
		2012	\$113k
		2011	\$52k

Table 10. Summary statistics of the distribution of the predicted insurance pure premium per household in British Columbia.

Quantile level	90%	95%	97.5%	99%	99.5%	100%
Pure premium	30	59	208	1,147	1,753	7,980

grid cells. This is because those cells have predicted annual burn probability of 0. [Table 10](#) summarizes the quantiles of the insurance pure premium per household over all grid cells.

6. DISCUSSION

Wildfire risk in North America is of great significance and over the past decade, insurers have been suffering annually from billions of dollars in losses. This article proposes a transparent and simple, yet powerful model that relies on random forests. It can be possibly used by actuaries to price, reserve or manage the financial risk from wildfires. Unlike wildfire models that were designed for a specific region, the models presented in this article cover a wide geographic area. They have high predictive capabilities with strong ability to classify locations into high/low risk zones. The random forest models are characterised by their ability to capture non-linearity and interaction in the model inputs. Occurrence of wildfire can be devastating, thus it is of utmost importance to measure wildfire risk by calculating the burn probabilities of wildfires in high-risk areas with densely populated communities.

Calculating insurance premiums is a very important task for actuaries, and it is a complex endeavor for emerging risks such as wildfires. The approach described in Section 5 along with the Leaflet maps provided (see Supplementary Material) can be used as a baseline description of the wild-

fire hazard. With extensive property characteristics data obtained at the underwriting stage and appropriate vulnerability curves, actuaries can therefore compute the loss costs necessary for premium determination and accurate underwriting. Moreover, some actuarial applications such as reserving or capital requirements might require a representation of spatial dependence. By construction, the wildfire models presented in Section 3 link occurrence to a set of covariates, and as such, borrows spatial dependence from e.g., temperature and precipitation dynamics. The actuary shall therefore make sure the inputs to the wildfire model are spatially consistent over the desired time horizon for the aforementioned applications.

SUPPLEMENTARY MATERIAL

Leaflet maps for the annual burn probabilities of wildfires of any size and of large wildfires are available in <https://robabairakdar.shinyapps.io/WildfireLikelihood/>, and <https://robabairakdar.shinyapps.io/LargeWildfireLikelihood/>, respectively. The maps are available for the three high-risk states and provinces: Alberta, British Columbia and California. They are desktop and mobile friendly interactive maps that provide the user with the flexibility to zoom on a certain location on a map. The leaflets can be used by insurers to obtain the burn probabilities for any location in the selected high-risk states and provinces.

ACKNOWLEDGEMENTS

This research was supported in part by the Casualty Actuarial Society. Roba Bairakdar recognizes the support of the Society of Actuaries' James C. Hickman Scholar Program. Mathieu Boudreault acknowledges the support of the Natural Sciences and Engineering Research Council of Canada RGPIN-2021-03362. Méлина Mailhot recognizes the support

of the Natural Sciences and Engineering Research Council of Canada RGPIN-2015-05447.

Submitted: May 11, 2022 EDT. Accepted: January 17, 2023 EDT. Published: July 31, 2023 EDT.

REFERENCES

- Atkinson, Dale, Mark Chladil, Volker Janssen, and Arko Lucieer. 2010. "Implementation of Quantitative Bushfire Risk Analysis in a GIS Environment." *International Journal of Wildland Fire* 19 (5): 649. <https://doi.org/10.1071/wf08185>.
- Breiman, Leo. 2001. *Machine Learning* 45 (1): 5–32. <https://doi.org/10.1023/a:1010933404324>.
- Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 2001. *Classification and Regression Trees*. Routledge.
- Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly Media, Inc.
- CALFIRE (California Fire Department of Forestry and Fire Protection). 2022. "Top 20 Largest California Wildfires." 2022. <https://www.fire.ca.gov/>.
- CatIQ (Canada's Loss and Exposure Indices Provider). 2021. 2021. <https://public.catiq.com>.
- CDS (Copernicus Climate Change Service (C3S) Climate Data Store). 2019. "ERA5-Land Hourly Data from 1950 to Present." Muñoz Sabater, J. and Nicolas, J. and Peubey, C. and Radu, R. and Rozum, I. and Schepers, D. and Simmons, A. and Soci, C. and Dee, D. and Thépaut, J-N. 2019.
- Cecil, Daniel J., Dennis E. Buechler, and Richard J. Blakeslee. 2014. "Gridded Lightning Climatology from TRMM-LIS and OTD: Dataset Description." *Atmospheric Research* 135–136 (January):404–14. <https://doi.org/10.1016/j.atmosres.2012.06.028>.
- Center For International Earth Science Information Network-CIESIN-Columbia University. 2017. "Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11." Palisades, NY: Socioeconomic Data and Applications Center (SEDAC). <https://doi.org/10.7927/H49C6VHW>.
- CFS (Canadian Forest Service). 2019. *National Fire Database - Agency Fire Data*. Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, Alberta. http://cwfis.cfs.nrcan.gc.ca/en_CA/nfdb.
- Chuvieco, Emilio, Inmaculada Aguado, Marta Yebra, Héctor Nieto, Javier Salas, M. Pilar Martín, Lara Vilar, et al. 2010. "Development of a Framework for Fire Risk Assessment Using Remote Sensing and Geographic Information System Technologies." *Ecological Modelling* 221 (1): 46–58. <https://doi.org/10.1016/j.ecolmodel.2008.11.017>.
- CIFFC (Canadian Interagency Forest Fire Centre). 2021. "National Fire Situation Report, Sep 15, 2021." 2021. <https://ciffc.net/en/ciffc/sitrep/2021-09-15>.
- Cragg, John G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica* 39 (5): 829. <https://doi.org/10.2307/1909582>.
- EM-DAT (Emergency Events Database). 2020. "EM-DAT: The International Disasters Database." 2020. <http://www.emdat.be/>.
- Finney, Mark A., Charles W. McHugh, Isaac C. Grenfell, Karin L. Riley, and Karen C. Short. 2011. "A Simulation of Probabilistic Wildfire Risk Components for the Continental United States." *Stochastic Environmental Research and Risk Assessment* 25 (7): 973–1000. <https://doi.org/10.1007/s00477-011-0462-z>.
- Giglio, Louis, Luigi Boschetti, David P. Roy, Michael L. Humber, and Christopher O. Justice. 2018. "The Collection 6 MODIS Burned Area Mapping Algorithm and Product." *Remote Sensing of Environment* 217 (November):72–85. <https://doi.org/10.1016/j.rse.2018.08.005>.
- Groom, Nichola. 2015. "Wildfires Push Californians to High-Risk Insurance Market." Reuters. Thomson Reuters. 2015. <https://www.reuters.com/article/us-usa-wildfires-insurance-insight-idUSKCN0RL0AG20150921>.
- Hanes, Chelene C., Xianli Wang, Piyush Jain, Marc-André Parisien, John M. Little, and Mike D. Flannigan. 2019. "Fire-Regime Changes in Canada over the Last Half Century." *Canadian Journal of Forest Research* 49 (3): 256–69. <https://doi.org/10.1139/cjfr-2018-0293>.
- Kuhn, Max. 2021. *Caret: Classification and Regression Training*. <https://CRAN.R-project.org/package=caret>.
- Massada, Avi Bar, Volker C. Radeloff, Susan I. Stewart, and Todd J. Hawbaker. 2009. "Wildfire Risk in the Wildland–Urban Interface: A Simulation Study in Northwestern Wisconsin." *Forest Ecology and Management* 258 (9): 1990–99. <https://doi.org/10.1016/j.foreco.2009.07.051>.
- McCullagh, Peter. 1984. "Generalized Linear Models." *European Journal of Operational Research* 16 (3): 285–92. [https://doi.org/10.1016/0377-2217\(84\)90282-0](https://doi.org/10.1016/0377-2217(84)90282-0).

Mitchell-Wallace, Kirsten, Matthew Jones, John Hillier, and Matthew Foote. 2017. *Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide*. John Wiley & Sons.

NALCMS (North American Land Change Monitoring System). 2020. *2015 Land Cover of North America at 30 Meters*. Commission for Environmental Cooperation. <http://www.cec.org/nalcms>.

NOAA (National Oceanic and Atmospheric Administration). 2016. "World Digital Elevation Model ETOPO5." European Environment Agency. 2016. <https://www.eea.europa.eu/data-and-maps/data/world-digital-elevation-model-etopo5>.

Shi, Peng, and Kun Shi. 2017. "Territorial Risk Classification Using Spatially Dependent Frequency-Severity Models." *ASTIN Bulletin* 47 (2): 437–65. <https://doi.org/10.1017/asb.2017.7>.

Short, Karen C. 2017. "Spatial Wildfire Occurrence Data for the United States, 1992-2015 [FPA_FOD_20170508] (4th Edition)." *Forest Service Research Data Archive*. Forest Service Research Data Archive. <https://doi.org/10.2737/rds-2013-0009.4>.

Stocks, B. J., J. A. Mason, J. B. Todd, E. M. Bosch, B. M. Wotton, B. D. Amiro, M. D. Flannigan, et al. 2002. "Large Forest Fires in Canada, 1959–1997." *Journal of Geophysical Research* 108 (D1): FFR-5. <https://doi.org/10.1029/2001jd000484>.

Van Wagtendonk, Jan W. 1996. "Use of a Deterministic Fire Growth Model to Test Fuel Treatments." In *Sierra Nevada Ecosystem Project, Final Report to Congress*, 2:1155–66.

Westerling, A. L., H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam. 2006. "Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity." *Science* 313 (5789): 940–43. <https://doi.org/10.1126/science.1128834>.

Yaghmaei, Nima. 2020. *Human Cost of Disasters: An Overview of the Last 20 Years, 2000-2019*. United Nations Office for Disaster Risk Reduction.