Supplement of Atmos. Chem. Phys., 20, 11065–11087, 2020
https://doi.org/10.5194/acp-20-11065-2020-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

Quantifying the effects of environmental factors on wildfire burned area in the south central US using integrated machine learning techniques

Sally S.-C. Wang and Yuxuan Wang

Correspondence to: Yuxuan Wang (ywang246@central.uh.edu)

Supplement

Table S1. Studies using statistical methods to estimate burned area

Region	period	Method	Spatial domain	Spatial scale (estimated; km^2)	Temporal scale	R^2	Reference
Canada	1959-1997	MLR	Ecoregions	466x466~1123x 1123	Monthly	36-64%	Flannigan et al. (2005)
Portugal	1980-2004	MLR	Portuguese districts	25x25~100x100	Monthly	43-80%	Carvalho et al. (2008)
Alaska and Canada	1960-2002	MARS	Alaska and western Canada	100x100~235x2 35	Annually	82%	Balshi et al. (2009)
EU-Mediterranean		MLR	European Mediterranean basin	1400x1400	Monthly	87%	Camia and Amatulli (2009)*
Western US	1916-2003	MLR [1]	Ecoregions	600x600~1000x 1000	Annually	25-57%	Listtell et al. (2009)
Western US	1980-2004	MLR	Ecoregions	600x600~1000x 1000	Annually	37-57%	Spracklen et al. (2009)
Western US	1972-1988; 1989-2005	MLR, RF [2]	NUTS3[4]	600x600~1000x 1000	Annually	73%; 83%	Westerling et al. (2011)
EU-Mediterranean	1985-2004	MLR	Countries	300x300~1000x 1000	Monthly	39-69%	Amatulli et al. (2013)
EU-Mediterranean	1985-2004	RF	Countries		Monthly	33-72%	Amatulli et al. (2013)
EU-Mediterranean	1985-2004	MARS [3]	Countries		Monthly	43-77%	Amatulli et al. (2013)
Spain	1990-2008	MARS	Phytoclimatic zones	25x25~100x100	Monthly	1-37%	Bedia et al. (2013)
Western US	1916-2004	MLR	Ecoregions	600x600~1000x 1000	Annually	25-60%	Yue et al. (2013)
Western US	1916-2004	Parameteriza tion	Ecoregions		Annually	1-69%	Yue et al. (2013)
North-eastern Spain	1983-2012	MLR	Catalonia	300x300	Annually	33%	Marcos et al. (2015)
Iberian Peninsula	1981-2005	MLR	Pyro-regions	200x200~700x7 00	Monthly	52-72%	Sousa et al. (2015)
EU-Mediterranean	1985-2011	MLR	EUMED [5]	2000x2000	Annually	60%	Urbieta et al. (2015)
Pacific western coast of USA	1985-2011	MLR	Oregon and California	1000x1000	Annually	37%	Urbieta et al. (2015)
British Columbia, Canada	1961-2010	MLR	Southern Cordillera	1400x1400	Annually	55%	Kirchmeier-Young et al. (2018)

Catalonia, Iberian Peninsula	1982-2015	MLR	Fire regime zones	80x80~150x150	Annually	57-91%	Duane et al. (2019)
South-Central US	2002-2015	Integration of RF, logistic regression, and QRF	Eastern Texas, Oklahoma, Louisiana, and Arkansas	700x700	Monthly	50% (winter-spring); 79% (summer)	This study

[1] MLR: Multiple Linear Regression; [2] RF: Random forest; [3] MARS: Multivariate adaptive regression Splines; [4] NUTS3: Nomenclature of Territorial Units at the third level [5] EUMED: burned area summation over Portugal, Spain, South France, Italy and Greece

*: focus on only large fires

Table S2. Comparison of MAE and skewness between the RF model and the developed model

Metrics	**Model developed in this study**	**MLR**	**RF alone**	**XGboost**
MAE (winter-spring)	1.13	1.44	1.34	1.26
Skewness (winter-spring) §	0.70 (percentiles)	37.40 (burned area)	37.40 (burned area)	37.40 (burned area)
MAE (summer)	0.57	0.76	0.70	0.67
Skewness (summer) §	0.96 (percentiles)	33.83 (burned area)	33.83 (burned area)	33.83 (burned area)

§: Skewness: The calculation of the skewness is described below in the section of calculation of skewness.

Table S3. The selected XGboost hyperparameters for the winter-spring and summer fire seasons

	eta	**Max_depth**	**gamma**	**subsample**	**Colsample_bytree**	**Min_child_weight**
Winter-spring	0.01	10	3	0.75	0.7	1
Summer	0.05	8	3	1	0.6	1

20

Table S4. Model performance at grid level for the selected years

Year/Fire season	Including misclassified grids			Excluding misclassified grids		
	R^2	**RMSE (km^2)**	**MAE (km^2)**	**R^2**	**RMSE (km^2)**	**MAE (km^2)**
2011 (combine winter-spring and summer)	0.30	11.04	3.38	0.42	21.06	5.25
2014 (winter-spring)	0.30	5.19	1.05	0.51	5.87	0.77
2008 (summer)	0.42	1.58	0.38	0.66	1.75	0.43

Table S5. Model performance at grid level for each year

	Year													
Statistics	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Winter-spring (excluding misclassified grids)														
R^2	0.70	0.76	0.77	0.64	0.38	0.52	0.51	0.38	0.54	0.36	0.56	0.55	0.41	0.61
MAE (km^2)	0.20	0.46	0.34	0.34	5.63	0.37	2.10	3.12	0.38	3.17	0.32	0.27	0.83	0.34
RMSE (km^2)	2.02	2.30	1.50	1.91	14.61	1.64	11.18	14.21	1.81	23.34	2.48	0.91	6.04	1.92
Summer (excluding misclassified grids)														
R^2	0.40	0.46	0.60	0.62	0.59	0.31	0.59	0.43	0.47	0.40	0.49	0.56	0.37	0.41
MAE (km^2)	0.08	0.17	0.09	0.52	0.93	0.02	0.42	0.58	0.33	3.71	1.20	0.32	0.15	0.68
RMSE (km^2)	0.32	1.88	0.92	2.08	2.48	0.09	1.42	6.37	0.95	12.08	9.23	1.81	1.66	4.35
Winter-spring (including misclassified grids)														
R^2	0.40	0.52	0.49	0.39	0.35	0.27	0.28	0.29	0.31	0.23	0.35	0.28	0.30	0.40
MAE (km^2)	0.18	0.35	0.26	0.31	4.84	0.33	1.92	2.83	0.311	2.78	0.25	0.24	1.05	0.35
RMSE (km^2)	1.70	1.95	1.23	1.78	13.42	1.59	10.26	13.24	1.55	21.12	2.10	0.81	5.19	3.21
Summer (including misclassified grids)														
R^2	0.28	0.10	0.28	0.33	0.45	0.10	0.42	0.29	0.31	0.39	0.32	0.40	0.20	0.26

MAE (km^2)	0.09	0.18	0.11	0.42	0.76	0.05	0.38	0.48	0.31	3.09	1.08	0.27	0.19	0.57
RMSE (km^2)	0.67	1.75	0.95	1.79	2.20	0.25	1.58	5.50	0.96	10.85	8.35	1.63	2.06	3.85

Table S6. The ratio of %IncMSE at variable ranked as X percentile (Y^{th}) to the %IncMSE at variable ranked as $(Y+1)^{th}$ for the three selected percentiles

	25^{th} percentile (Y=14)	**50^{th} percentile (Y=29)**	**75^{th} percentile (Y=43)**
Spring-winter	1.21	0.88	1.00
Summer	1.06	1.01	1.00

Table S7. Comparison of accuracy, AUC, F-1 score, MAE, RMSE, and MAE of large burned area between the model with the chosen set of predictors and the model with the predictors that have lower degrees of collinearity ($r < |0.5|$)

Model	**Model with the chosen set of variables**	**Model with variables that have lower degrees of collinearity**
Winter-Spring		
Number of predictor variables	58	33
Accuracy (winter-spring)	0.74	0.71
AUC (winter-spring)	0.82	0.78
F-1 (winter-spring)	0.79	0.77
MAE (log(area); winter-spring)	1.37	1.43
RMSE (log(area); winter-spring)	2.03	2.06
MAE of large burned area† (log(area); winter-spring)	2.32	2.57
Summer		
Number of predictor variables	57	31
Accuracy (summer)	0.74	0.72
AUC (summer)	0.83	0.80
F-1 (summer)	0.77	0.75
MAE (log(area); summer)	1.17	1.20
RMSE (log(area); summer)	1.87	1.88

MAE of large burned area† (log(area); summer)	2.25	2.31

† Large burned area here is defined as the burned area larger than 90th percentile.

Table S8. Mean scaled absolute percentage effects by control factor group for the two fire seasons calculated by decompose analysis at the large-scale domain.

	Weather effect	**Fuel effect**	**Climate effect**	**Fix effect**	**Interaction effect**
Spring-winter	12.57	21.39	33.23	22.93	10.87
Summer	16.26	17.29	35.79	21.8	8.87

Table S9. The mean variable importance metrics (%IncMSE) of each effect for the two fire seasons calculated based on grid burned area prediction

	Weather effect	**Fuel effect**	**Climate effect**	**Fix effect**
Spring-winter	8.20	9.04	12.09	6.56
Summer	9.12	4.85	19.18	4.59

Table S10. Comparison of MAE, MAE of large burned area, and standard deviation of predictions between the model with the chosen percentiles, percentile test set 1, and percentile test set 2.

Model	**With the chosen percentiles***	**Percentile test set 1***	**Percentile test set 2***
MAE (log(area); winter-spring)	1.37	1.30	1.29
MAE of large burned area† (log(area); winter-spring)	1.97	2.64	2.81
Standard deviation of predictions (log(area); winter-spring)	2.42	2.09	2.09
MAE (log(area); summer)	1.17	1.12	1.11

MAE of large burned area† (log(area); summer)	2.25	2.42	2.52
Standard deviation of predictions (log(area); summer)	2.19	1.93	1.92

* Model developed in this study: Use the selected percentiles of 45, 55, 65, 85, 95, and 99 and six subgroups of (39, 49), (50, 59), (60, 69), (70, 79), (80, 89), (>=90).

* Set 1: Use the selected percentiles of 45, 55, 65, 75, 85, and 95 and six subgroups of (39, 49), (50, 59), (60, 69), (70, 79), (80, 89), (>=90).

* Set 2: Use the selected percentiles of 47.5, 63, 78, and 93 and four subgroups of (39, 55), (56, 70), (71, 85), (86, 100).

† Large burned area here is defined as the burned area larger than 90th percentile.

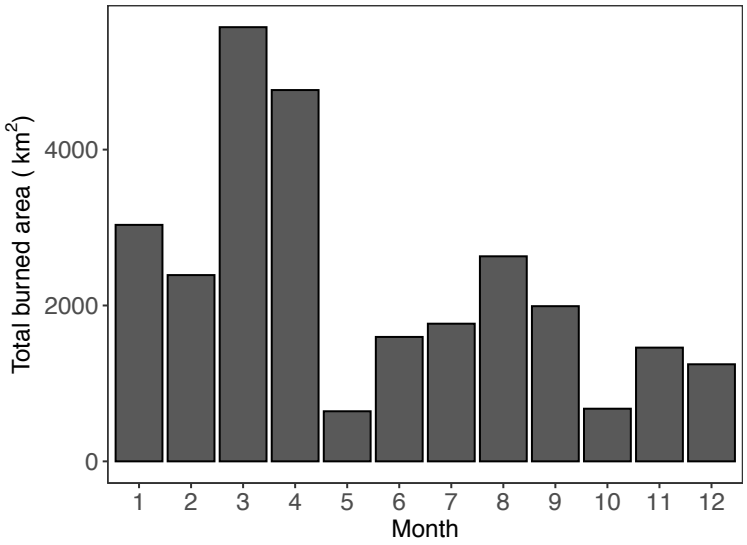


Figure S1. Seasonal burned area for the South Central US. It shows the monthly total burned area summed over 2002-2015.

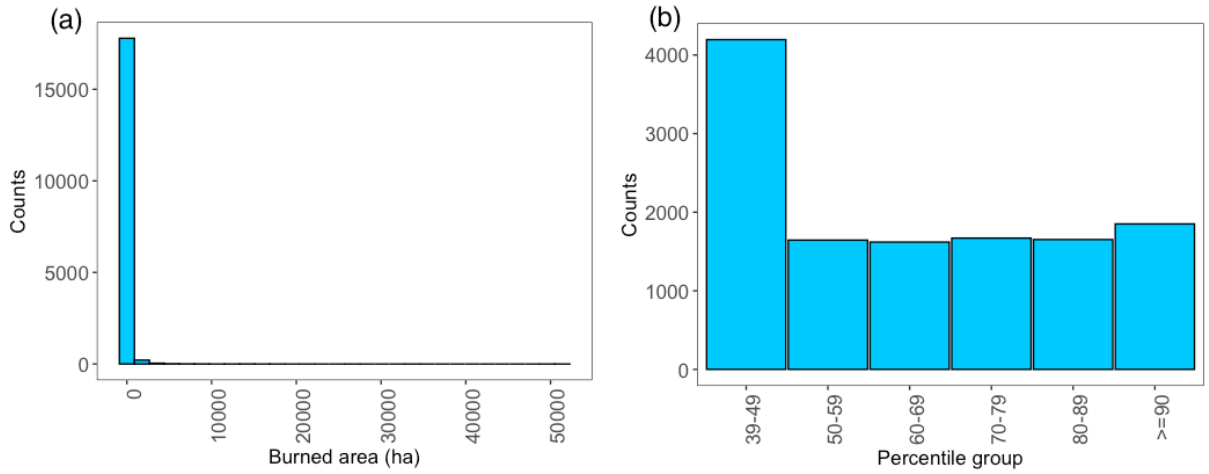


Figure S2. (a) Histogram of burned area of all the grids and (b) histogram of the percentile groups of burned area for the winter-spring fire season.

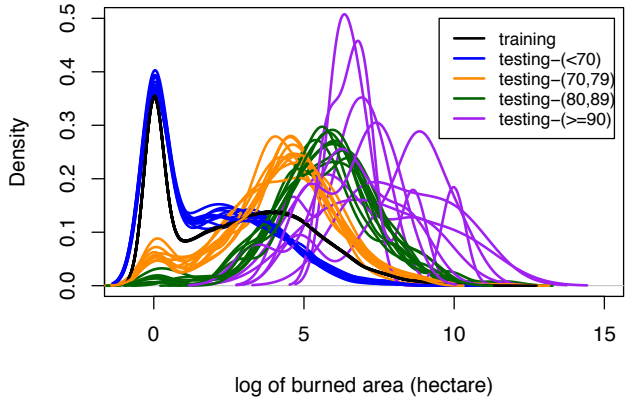


Figure S3. Probability distribution of burned area for 10 folds of the training set (black line), testing set predicted to have percentiles less than 70 (blue), between 70 and 79 (yellow), between 80 and 89 (green), and equal to or larger than 90 (purple).

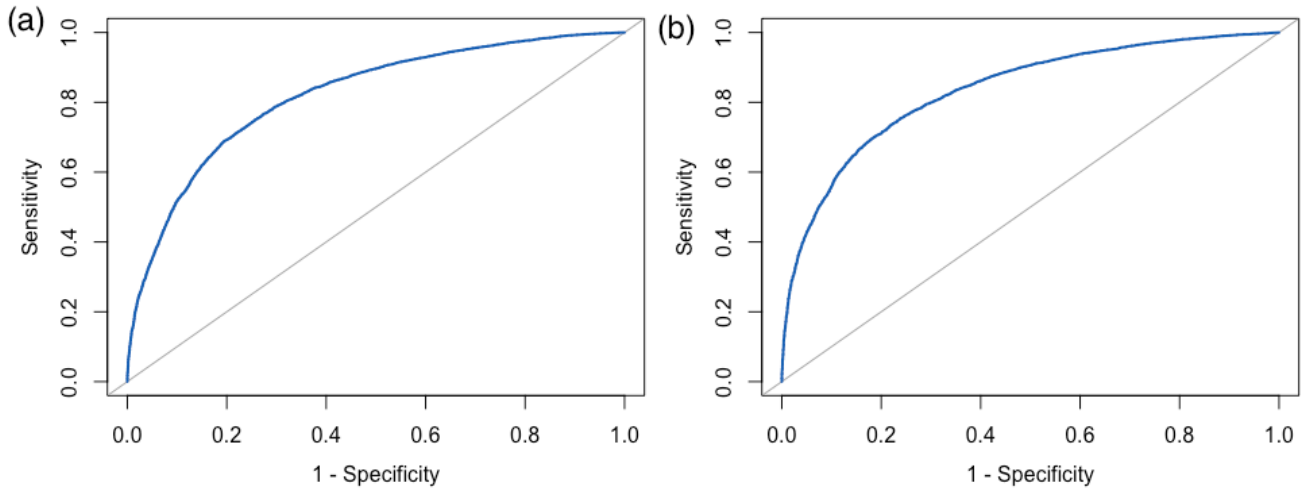


Figure S4. The ROC curve analysis of the logistic model for predicting burned grids in (a) winter-spring and (b) summer fire season.

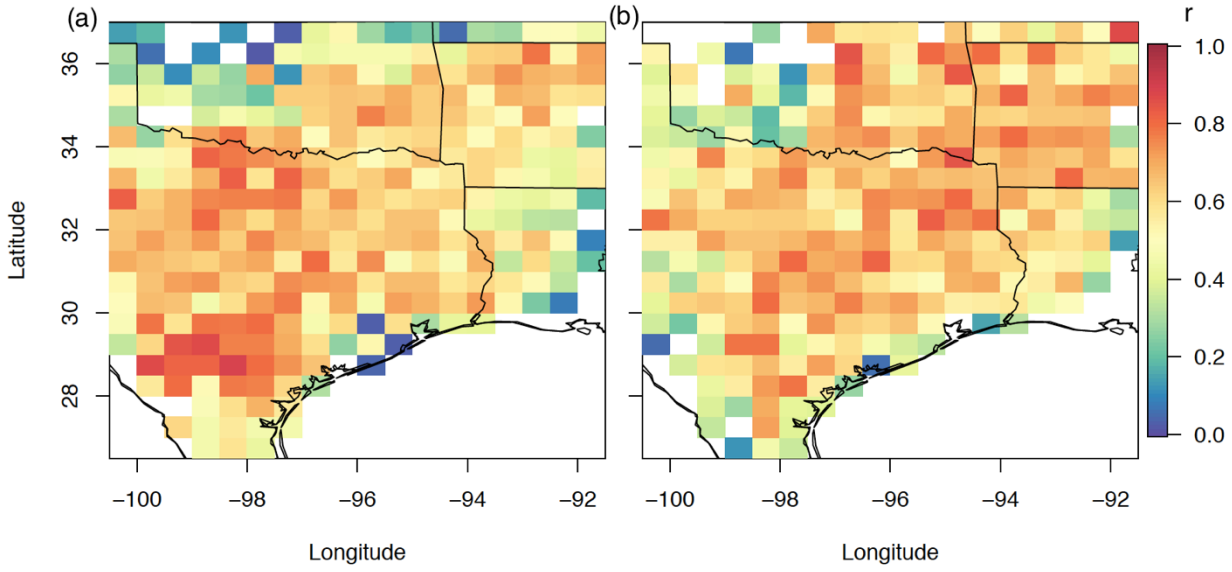


Figure S5. Maps of temporal correlation between observed and predicted burned area for each grid for the (a) winter-spring fire season (b) summer fire season.

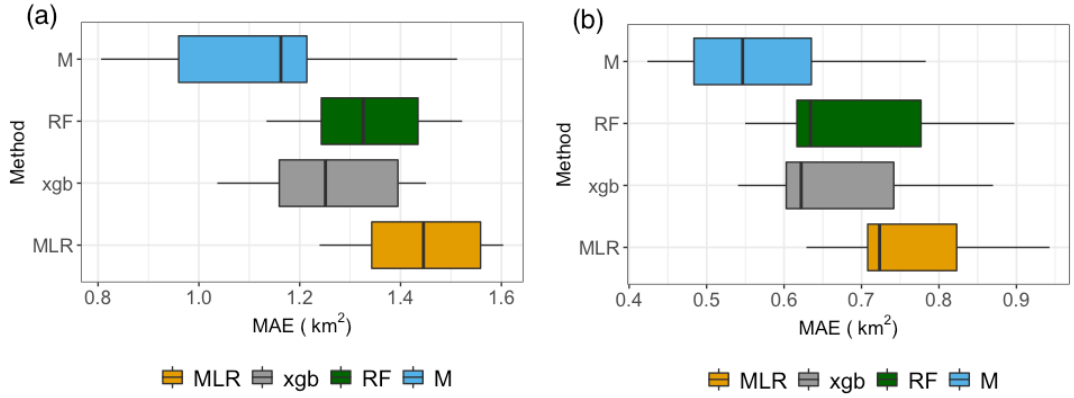


Figure. S6. Box plots of MAE from 10-fold-cross validation and different methods for (a) winter-spring and (b) summer fire season.

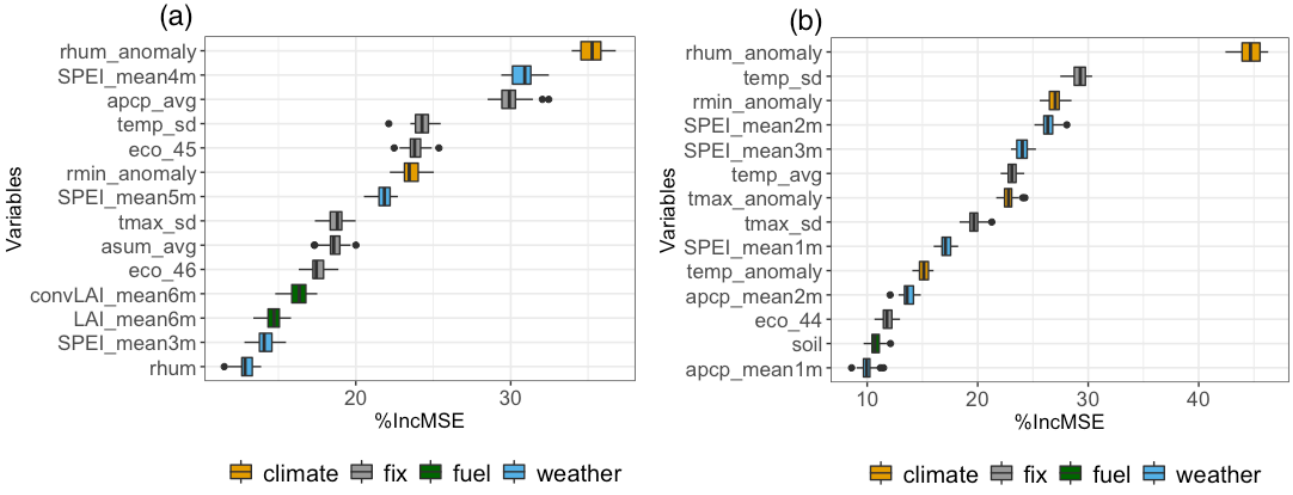


Figure S7. Box plots of variable importance in %IncMSE from the 50 times 10-fold cross validation for (a) winter-spring and (b) summer fire season.

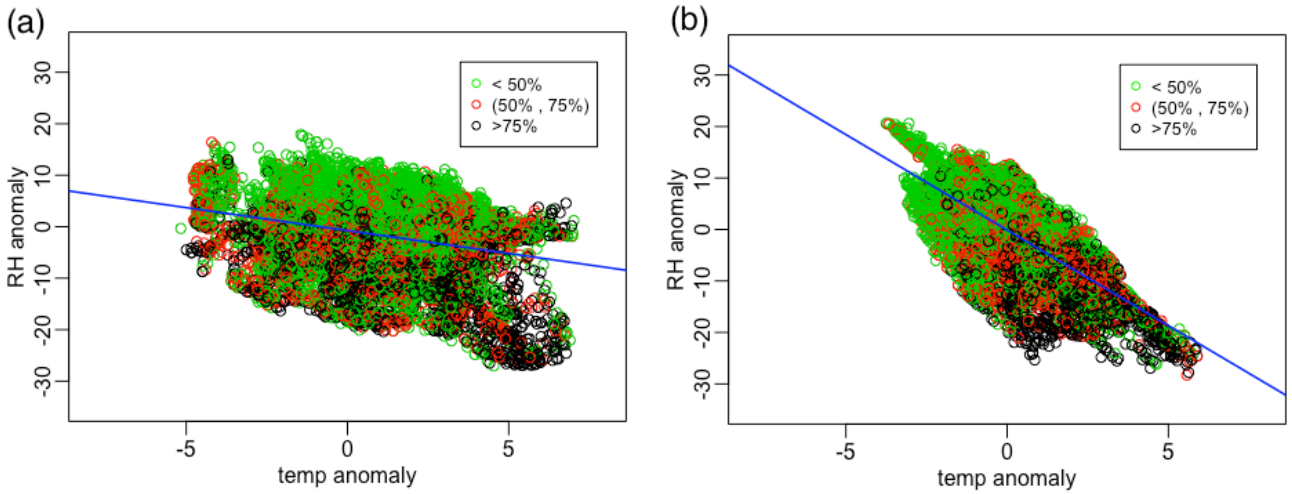


Figure S8. Scatter plot of RH anomaly versus temperature anomaly for (a) winter-spring and (b) summer fire season. The color represents different sizes of fire burned area (Green: smaller than 50th percentile; Red: larger than 50th percentile but smaller than 75th percentile; Black: larger than 75th percentile).

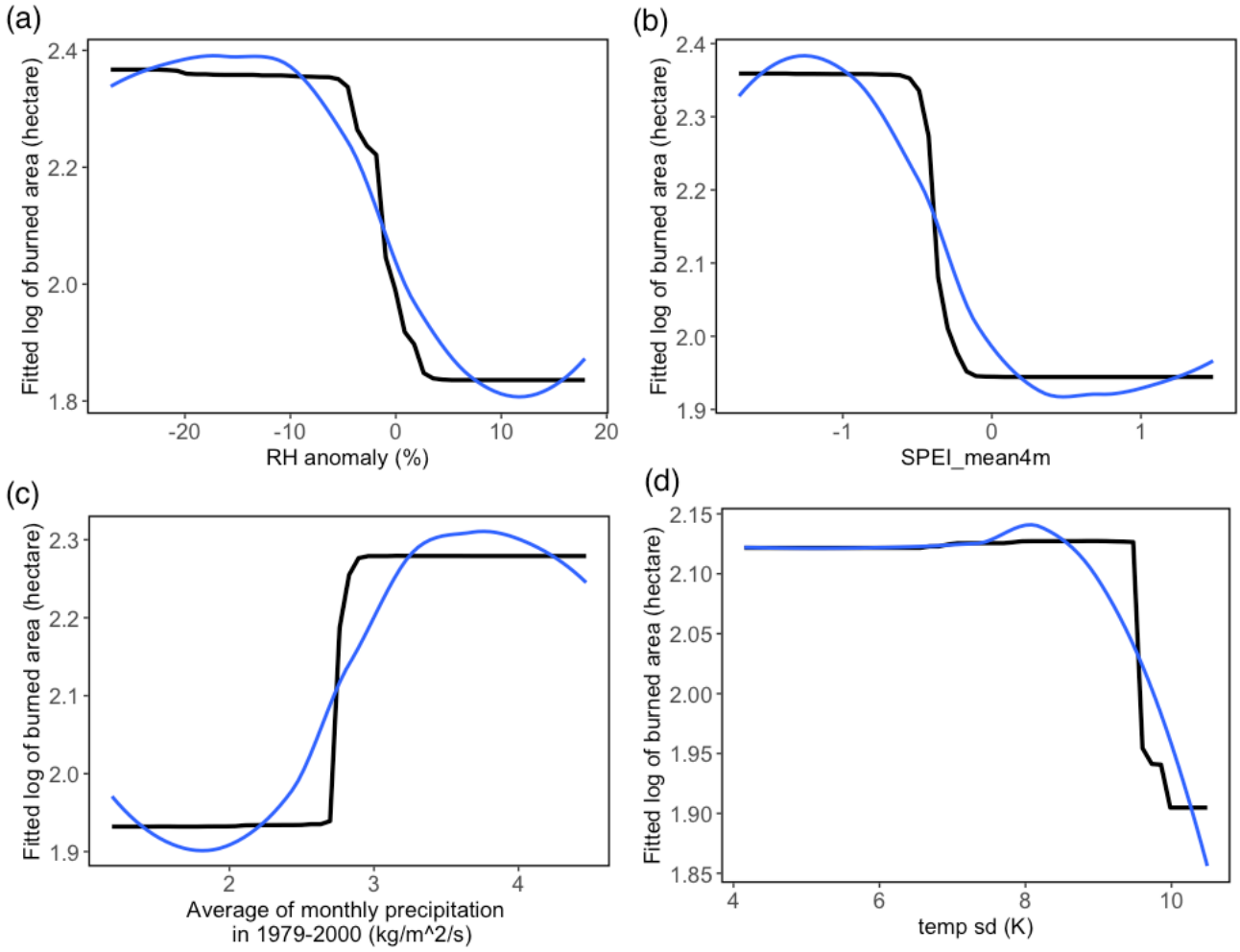


Figure S9. Partial dependence plots for the burned area model and (a) RH anomaly, (b) the mean SPEI of the preceding 4 months, (c) the average precipitation of 1979-2000, (d) the standard deviation of temperature of 1979-2000 for the winter-spring fire season. The blue line is the LOESS smooth line.

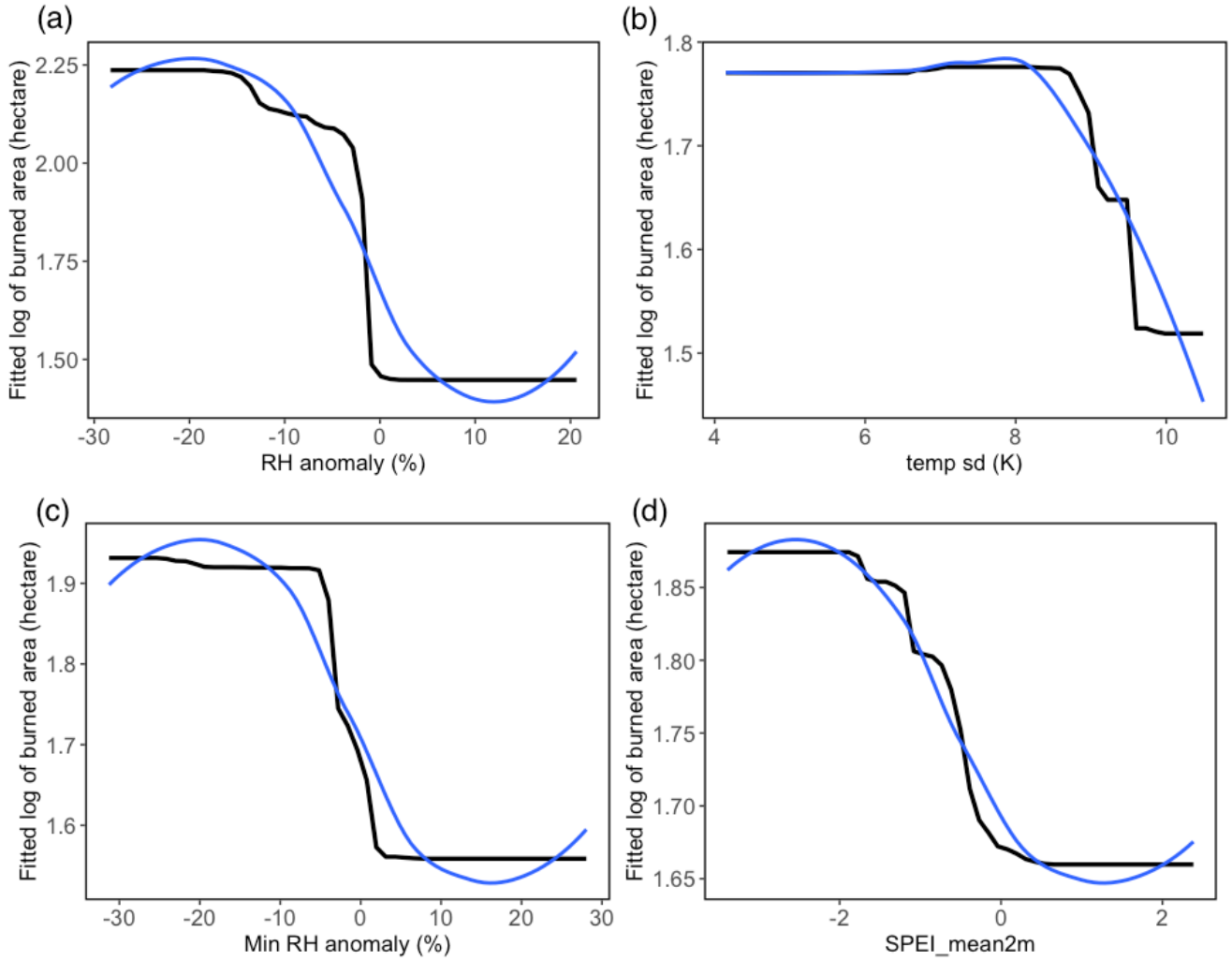


Figure S10. Partial dependence plots for the burned area model and (a) RH anomaly, (b) long-term (1979-2000) standard deviation of temperature, (c) minimum RH anomaly, and (d) the mean SPEI of the preceding 2 months for the summer season. The blue line is the LOESS smooth line. The blue line is the LOESS smooth line.

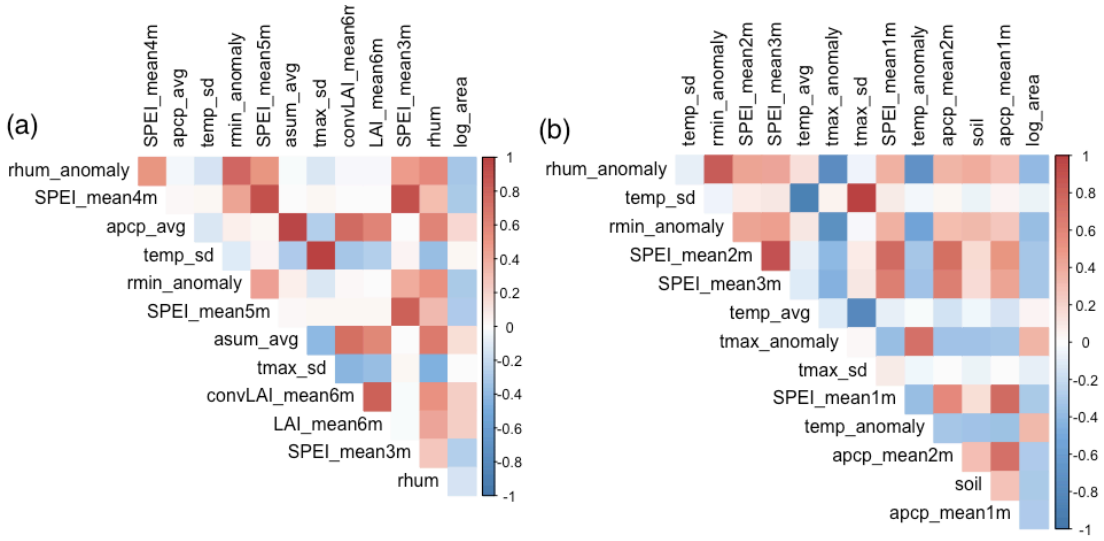
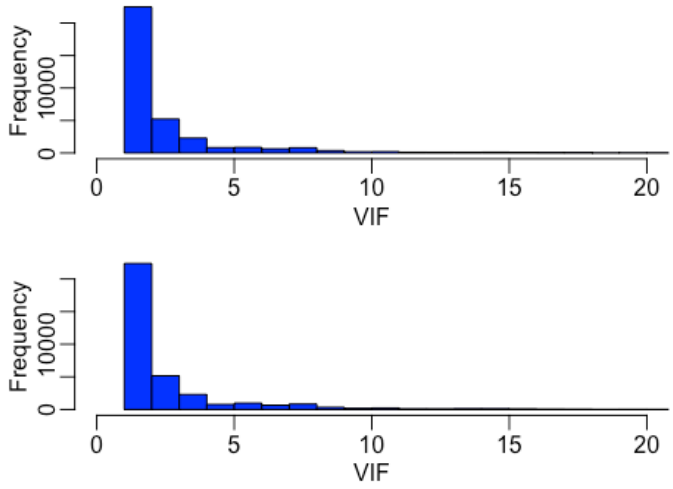
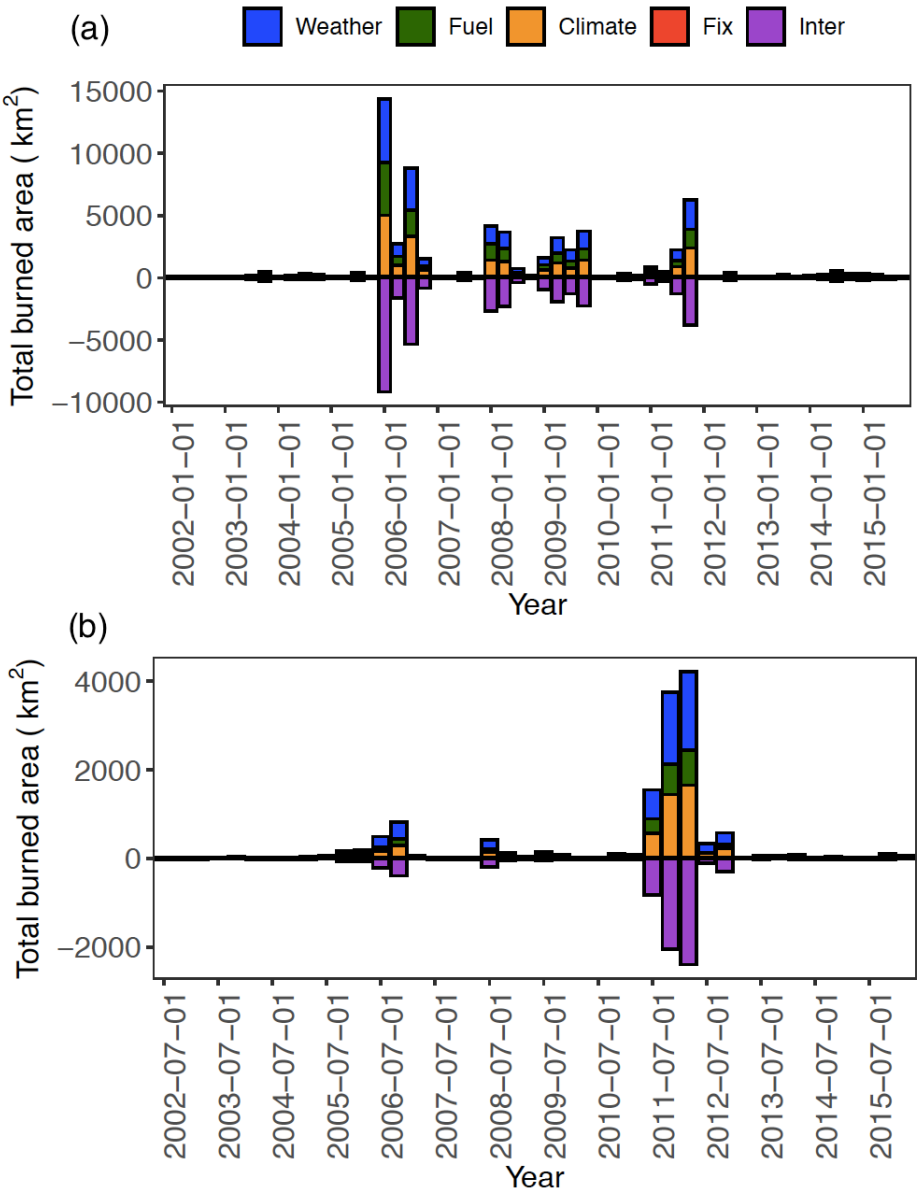


Figure S11. The correlation plot of the top 14 variables for the (a) winter-spring and (b) summer fire season.



105

Figure S12. Distributions of VIF calculated based on randomly selected seven variables of 5000 times sampling for winter-spring (top) and summer fire season (bottom).



110 **Figure S13.** Timeseries of burned area contributed by different environmental controls for the (a) winter-spring and (b) summer fire season. Color of blue, green, yellow, red, and purple indicate effect of weather, fuel, climate, fix, and interaction.

115

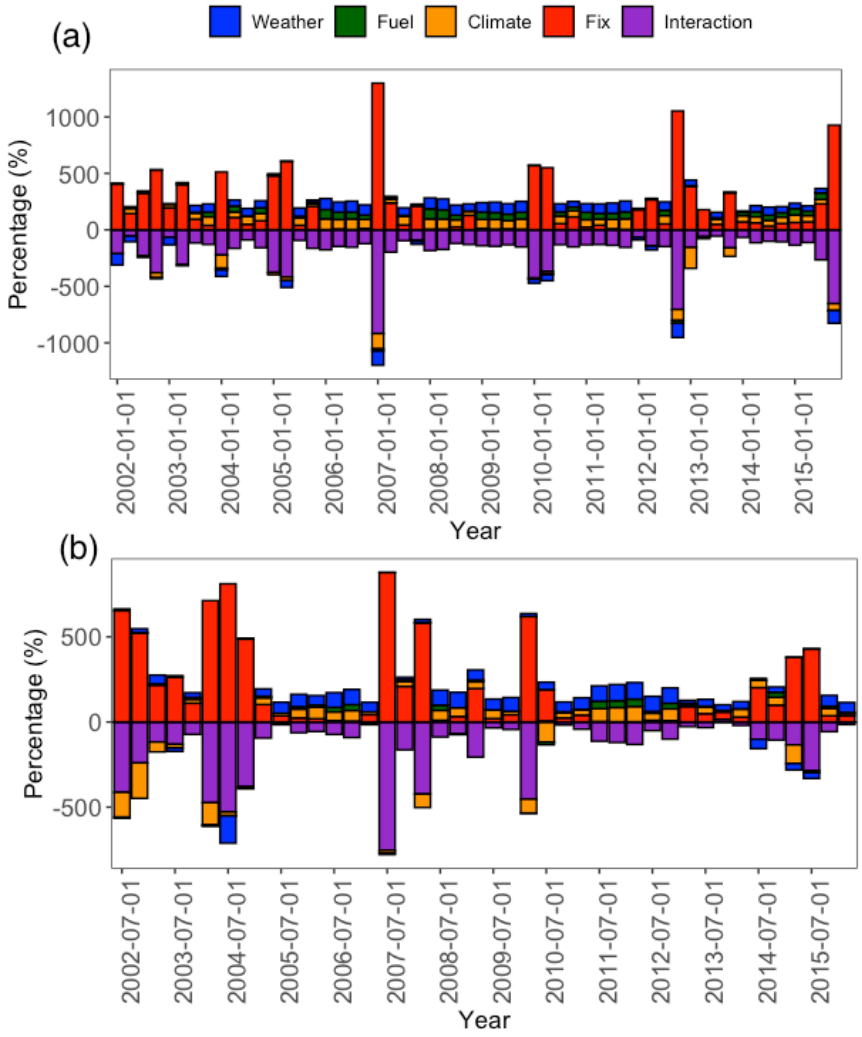
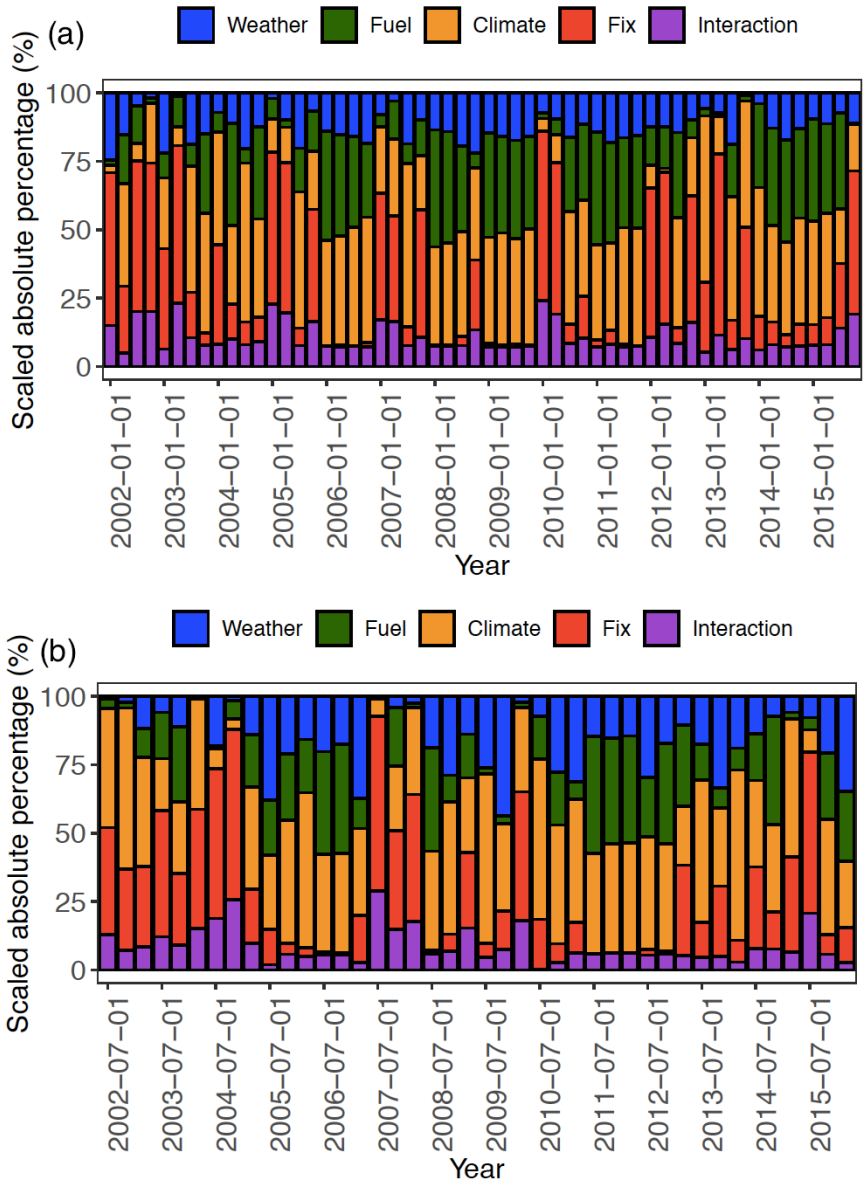


Figure S14. Timeseries of the percentage for the (a) winter-spring fire season and (b) summer fire season. Color of blue, green, yellow, red, and purple indicate effect of weather, fuel, climate, fix, and interaction. The percentage was calculated by dividing the total burned area of the month.



125

Figure S15. Timeseries of the scaled absolute percentage for the (a) winter-spring fire season and (b) summer fire season. Color of blue, green, yellow, red, and purple indicate effect of weather, fuel, climate, fix, and interaction.

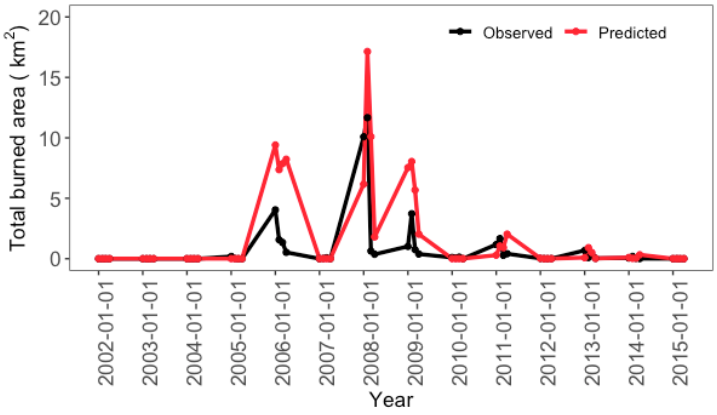


Figure S16. Timeseries of observed (black line) and predicted total burned area (red line) for the selected grid (Lon: -98.75, Lat: 29.25) for the winter-spring fire season.

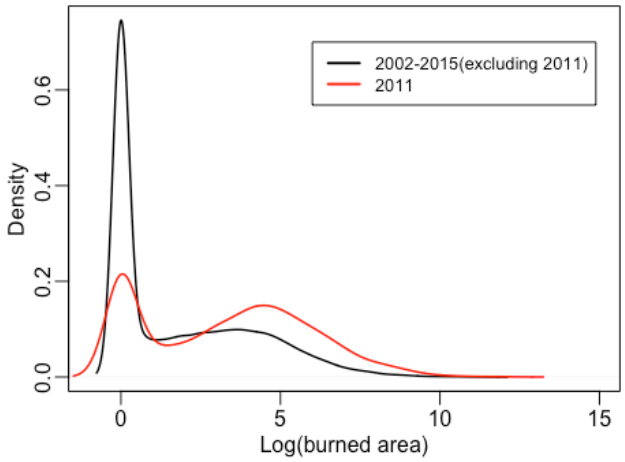


Figure S17. Distribution of burned area of all the grids for the study period excluding 2011 (black line) and of the grids for the extreme year 2011 (red line) combined both seasons.

XGBoost Model

XGBoost (eXtreme Gradient Boosting) is based on Gradient Boosting Decision Tree (GBDT) method, which is an iterative decision tree algorithm. GBDT iterates multiple trees to make final decisions. Compared to GBDT, XGBoost uses a more regularized model formalization to control over-fitting and parallelizes the tree formation to enhance the computational power (Chen and Guestrin, 2016). The hyperparameters of XGboost were tuned by a grid search with 10-fold cross-validation

to find the best model based on MAE. Table S3 shows the optimum value of each hyperparameter for the winter-spring and summer fire season model.

Calculation of skewness

Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. The skewness of a random variable X is the third standardized moment $\widetilde{\mu_3}$, defined as:

$$\widetilde{\mu_3} = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}} \quad (1)$$

where μ is the mean, σ is the standard deviation, E is the expectation operator, μ_3 is the third central moment, and κ_t are the *t*-th cumulants. If skewness is less than -1 or greater than +1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between +0.5 and +1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric. The positive value indicates that the tail is on the right side of the distribution while negative value indicates that the tail is on the left.

Calculation of correlation coefficient (r)

We also calculated two types of correlation coefficient (r) to evaluate model performance: spatial r and temporal r. For the spatial r, for each month, we calculated the correlation coefficient of the prediction and observation for all the grids over the whole domain. As for the temporal r, for each grid, we calculated the correlation coefficient for the timeseries of observed and predicted burned area. At the end, we obtain a map showing the temporal R, demonstrated in Figure S5.

Method to decompose the relative influence of environmental controls

A set of sensitivity experiments was designed to decompose the effect of different environmental controls across our study domain by perturbing variables belonging to one category at a time. The environmental control categories to be perturbed include weather, climate, and fuel. The fix-geospatial factors remain unchanged in each sensitivity experiment. The variables of each category are listed in Table 1. First, to examine the influence of weather, for each grid, we assigned the values of individual weather variables to their 15-year means by grids while keeping the variation of other variables (hereafter refer to as the "weather-avg run"). The same procedure was applied to the variables in the climate and fuel category, resulting in the climate-avg run and fuel-avg run respectively. The original model with all the variables of each grid varying by time is called the full-model run. Second, the gridded burned area predicted from each run is summed over all the grids across the study domain. The differences in resulting total burned area between the full-model run and weather-avg run represent the impact of weather control (hereafter called "weather effect"), and the same procedure was applied to derive the climate effect and fuel effect on the burned area. We also conducted the fixed run, in which for each grid, its weather, climate anomaly, and fuel variables are all fixed to their long-term mean, and the predicted burned area from this run represents the influence of geospatial

variables and climate normals on the burned area (hereafter named "fix effect"). Although the calculations of deriving the effect of a given environmental category are made by assuming linearity, the machine-learning-based prediction model does not assume linearity. Thus, the summation of burned area prediction from the weather, climate, fuel, and fixed run is not necessarily equal to the burned area predicted by the full model. This difference is considered as the interaction effect among these environmental controls.

After deriving the effects of the environmental controls on the burned area, we then calculated such effects of environmental controls in the scaled absolute percentage. The effect of an environmental control category was normalized by the number of variables in that category because the numbers of variables are different by environmental control and the category with a larger number of variables may have a larger effect on the burned area. Then, the scaled absolute percentage is defined as the normalized absolute value of the effect of one environmental control divided by the summation of the normalized absolute values of all the effects over all the categories. Thus, the scaled absolute percentage represents the average effect of a single variable in each category. For example, Equation (1) shows how we calculated the scaled absolute percentage of the weather contribution on burned area:

$$\frac{|E_w|/N_w}{|E_w|/N_w + |E_{fu}|/N_{fu} + |E_c|/Nc + |E_{fi}|/N_{fi} + |E_i|/N_t}, \quad (2)$$

, where E is the influence of the environmental controls in burned area, N indicates the number of variables in the category, N_t is the total number of variables, and the subscript *w, fu, c, fi, and, i* represent weather, fuel, climate, fixed, and interaction, respectively.