



Wildfire Susceptibility Assessment in Southern China: A Comparison of Multiple Methods

Yinxue Cao¹ · Ming Wang¹ · Kai Liu¹

Published online: 22 June 2017

© The Author(s) 2017. This article is an open access publication

Abstract Wildfire is a primary forest disturbance. A better understanding of wildfire susceptibility and its dominant influencing factors is crucial for regional wildfire risk management. This study performed a wildfire susceptibility assessment using multiple methods, including logistic regression, probit regression, an artificial neural network, and a random forest (RF) algorithm. Yunnan Province, China was used as a case study area. We investigated the sample ratio of ignition and nonignition data to avoid misleading results due to the overwhelming number of nonignition samples in the models. To compare model performance and the importance of variables among the models, the area under the curve of the receiver operating characteristic plot was used as an indicator. The results show that a cost-sensitive RF had the highest accuracy (88.47%) for all samples, and 94.23% accuracy for ignition prediction. The identified main factors that influence Yunnan wildfire occurrence were forest coverage ratio, month, season, surface roughness, 10 days minimum of the 6 h maximum humidity, and 10 days maxima of the 6 h average and maximum temperatures. These seven variables made the greatest contributions to regional wildfire susceptibility. Susceptibility maps developed from the models provide information regarding the spatial variation of ignition susceptibility, which can be used in regional wildfire risk management.

Keywords China · Random forest · Variable importance rank · Wildfire susceptibility · Yunnan forest

1 Introduction

As global temperatures warm, wildfires in China have become a significant concern because of their increasing frequency and severity, which has expanded the area that is affected by wildfires. Understanding wildfire susceptibility—defined as the likelihood of suffering harm—and its dominant influencing factors at a regional scale is necessary for improving wildfire management. The precision of wildfire forecasts needs to be improved, and understanding the driving forces of wildfires is of great importance for devising better strategies to mitigate wildfires and to identify at-risk areas (Finney 2005).

Wildfire susceptibility analyses are useful for fire occurrence prediction, zonation, and follow-up management. Statistical models, such as regression models, which have been developed from historical data to estimate the probabilities of fire occurrence under various local environmental conditions, are valuable for understanding general historical trends. These models can be used to predict outcomes, such as the expected number of fires in an area, from explanatory variables such as vegetation patterns, landforms, meteorological factors, and past fire history (Weinstein and Woodbury 2010).

Despite an abundance of studies that predict wildfire occurrence, many of these studies have employed traditional generalized linear models (GLMs), whose prediction accuracy is relatively low. In recent years, machine-learning methods have drawn researchers' attention because of their high modeling precision. Artificial neural networks (ANNs) are widely accepted machine-learning

✉ Ming Wang
wangming@bnu.edu.cn

¹ State Key Laboratory of Earth Surface Processes and Resource Ecology/Academy of Disaster Reduction and Emergency Management, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

methods, and they are often used as references with which to evaluate the performance of other machine-learning methods. ANNs are inspired by the sophisticated functionality of the human brain, where hundreds of billions of interconnected neurons process information in parallel, and researchers have successfully demonstrated that ANNs possess certain levels of intelligence (Hagan et al. 1996; Wang 2003). ANNs are capable of identifying complex, nonlinear relationships between input and output datasets. They have been used widely, particularly to address problems in which the characteristics of the underlying processes are difficult to describe using physical equations (Hsu et al. 1995).

Random forest (RF) algorithms are more recently developed, but they are some of the most mature machine-learning algorithms. RF algorithms have been used widely in data mining, bioinformatics, management, economics, and the medical sciences (Fang and Jian-Bina 2011). RF algorithms are ensemble classifiers that use decision trees as base classifiers, as proposed by Breiman (2001). They are based on Breiman's previous ensemble classifier bagging predictor (Breiman 1996), and they add more randomness by randomly selecting a subset of explanatory variables at each split node. This improves model accuracy and robustness simultaneously. Because of their high classification precision and stability, RF algorithms have been widely applied in economics, bioinformatics, environmental modeling of earthquakes or landslide susceptibilities (Catani et al. 2013), and forest fire susceptibility, and most of these algorithms have achieved better results than other methods (Aldersley et al. 2011; Oliveira et al. 2012; Massada et al. 2013; Rodrigues and de la Riva 2014).

Considering the extreme data imbalance regarding wildfire ignitions and nonignitions, a cost-sensitivity analysis is usually used in the models. Unlike most standard classifier learning algorithms that assume a relatively balanced class distribution and equal misclassification costs, a cost-sensitivity analysis ranks the importance of the classes and assigns different misclassification errors different penalties (Domingos 1999; Elkan 2001; Sun et al. 2007). This method resolves a great deal of the poor model classification performance that results from data imbalances (Del Río et al. 2014).

In this study, we established wildfire susceptibility models using logistic regression, probit regression, an ANN, a RF algorithm (RF-original), and a cost-sensitive RF algorithm (RF-cost sensitive). All of the models were fitted using the same explanatory and dependent variables, and they used the same randomly selected training samples. To evaluate model performance, we compared the five models in terms of their prediction accuracy, the importance of the explanatory variables, and their predicted spatial patterns of wildfire susceptibility.

2 Study Area and Data

Yunnan Province is located in southwestern China between $97^{\circ}31'–106^{\circ}11'E$ and $21^{\circ}08'–29^{\circ}15'N$. It belongs to the plateau type of the tropical monsoon climate zone, with cool summers and warm winters, and a mean annual temperature of $16^{\circ}C$. Yunnan has distinct dry and wet seasons, with a significantly uneven annual precipitation distribution. Winter and spring account for approximately 20% of the 1100 mm annual precipitation. Yunnan has three drought seasons: January to March, which affects two-thirds of the province; November to December, which affects one-half of the province; and April to early June, which affects 22% of the province (Peng et al. 2009). The continuous winter to spring drought is one of the main causes of wildfires in Yunnan (Chen et al. 2012). Since 2000, droughts from September to December and January to March have been more severe compared with historical records, which has led to a more frequent and intense winter to spring drought and increased wildfires in Yunnan. The occurrence of droughts of greater magnitude and longer duration suggests that megadroughts may occur in parts of Asia (including Yunnan) that are affected by the tropical monsoon. Such megadroughts can increase the risk of wildfires, as well as their scope and severity.

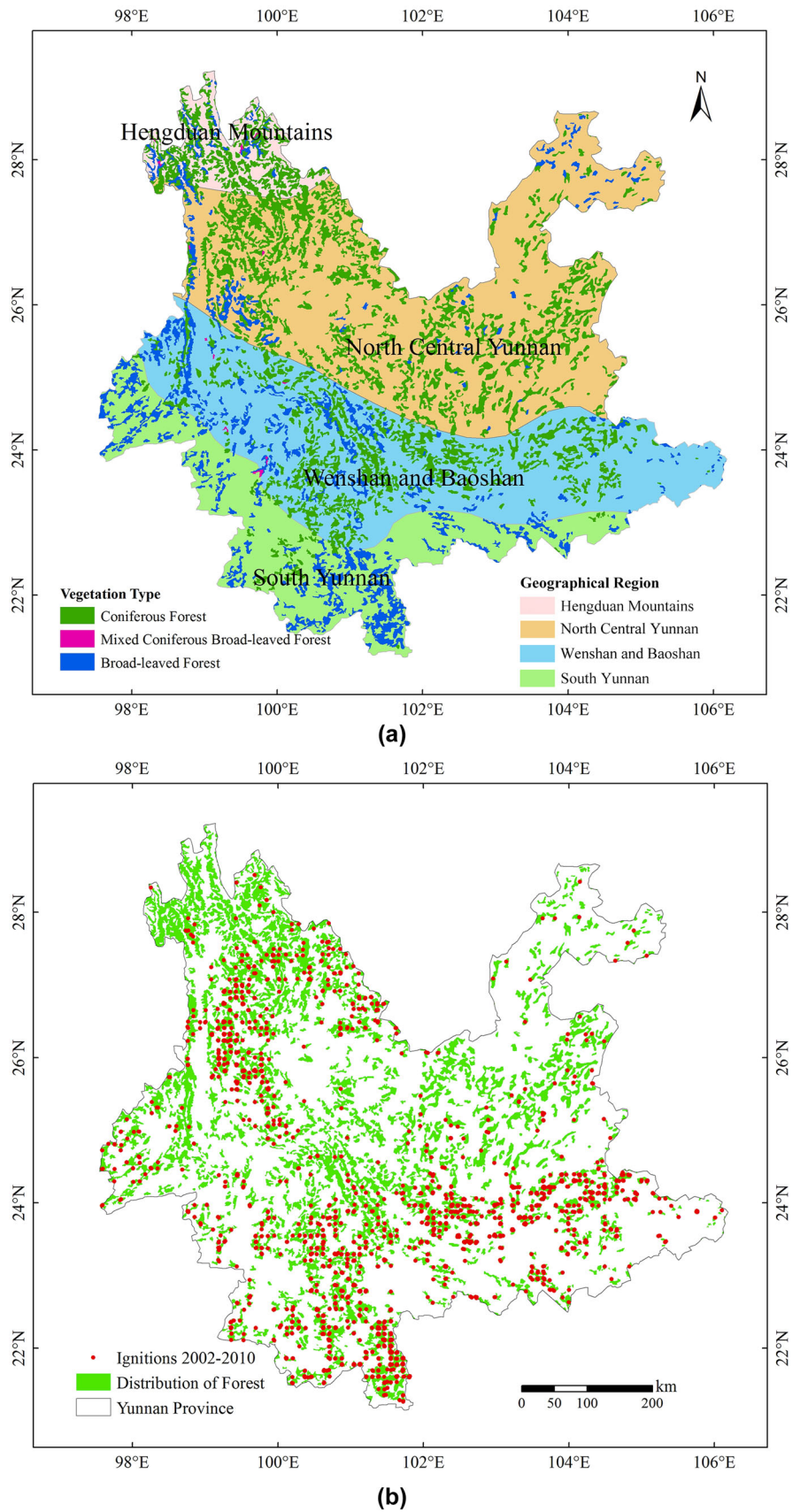
The forests in Yunnan include tropical and subtropical evergreen broad-leaved forests and temperate coniferous forests growing in cold mountain areas (Fig. 1a). Except for broad-leaved forests in southwestern Yunnan, coniferous forests are the dominant tree species in most areas. Wildfires in Yunnan mainly occur in winter and spring from December to May, and they are concentrated in spring from mid-February to mid-May (Chen et al. 2014).

The forest data in Yunnan used in this research were derived from the *Vegetation Map of the People's Republic of China (1:1,000,000)* (Zhang 2007). The wildfire ignition data in Yunnan used in this research were obtained from the 9-year time series of the maps of national thermal source distribution¹ from the National Satellite Meteorological Center of China with an interval of 10 days, from 2002 to 2010, with 324 images in total. The wildfire ignition vector points were gained by overlaying the boundary of Yunnan upon the forest distribution data. The distribution of the wildfire ignitions is shown in Fig. 1b. To demonstrate the difference between wildfires in different areas in Yunnan, we divided the study area into four parts, according to their integrated physical geographic characteristics (Huang 1989) as shown in Fig. 1a.

Figure 2 shows that wildfire ignitions were concentrated in winter and spring, and that very few wildfires occurred in summer and autumn. There was a gradually increasing

¹ <http://satellite.nsmc.org.cn/PortalSite/Default.aspx>.

Fig. 1 Distribution of forest in Yunnan Province, overlaid with the physico-geographical characteristics of the regions (a) and the wildfire ignition distribution, 2002–2010 (b). Data source a Huang (1989) and Zhang (2007); b <http://satellite.nsmc.org.cn/PortalSite/Default.aspx>



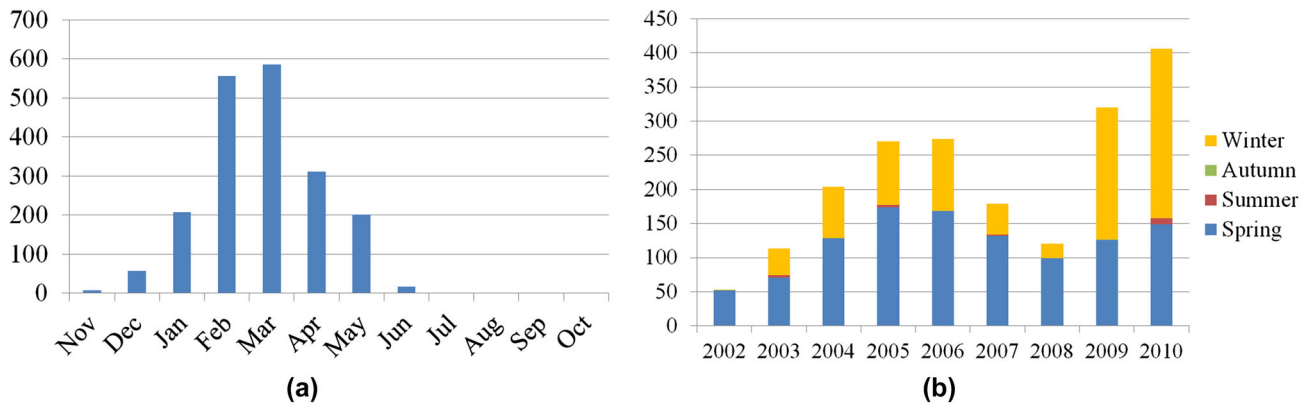


Fig. 2 Monthly (a) and seasonal and yearly (b) variations of the wildfire ignition number in Yunnan Province

yearly trend. From 2002 to 2005, the number of wildfire ignitions increased and this trend peaked in 2005 when 274 ignitions occurred. In 2006, the number of wildfire ignitions (273) was almost identical to that in 2005, but total ignitions decreased in 2007 and 2008. In 2009 and 2010, the number of wildfire ignitions increased dramatically (321 in 2009 and 406 in 2010), with a significant increase in winter wildfires. Satellite-detected fire ignitions included medium and large fires, which were the focus of this research, while the large number of small fires (usually less than 1 ha) were not considered. In addition, we also obtained the wildfire ignition statistical data for Yunnan Province between 1990 and 2015 from the *China Forestry Statistical Yearbook* (State Forestry Administration 1990–2015).

Three categories of variables that can potentially influence the susceptibility of wildfire ignitions were considered. They include meteorologically-related variables, vegetation-related variables, and landform-related variables. The meteorologically-related variables, which are the most important explanatory variables, were obtained from the climate forecast system reanalysis (CFSR) (Saha et al. 2010). The CFSR is the latest global reanalysis produced by the National Centers of Environmental Prediction (NCEP)² of the United States. As the latest generation of reanalysis data, the atmospheric-, oceanic-, and land surface-analyzed products of the CFSR have some of the highest horizontal resolutions, which reach approximately $0.312^\circ \times 0.312^\circ$ from 0°E to 359.688°E and 89.761°N to 89.761°S (1152 grids \times 576 grids, longitude/Gaussian latitude), with a 6 h resolution. Compared with other reanalysis data, such as ERA-40,³ JRA,⁴ MERRA,⁵ and

NCEP/NCAR,⁶ the CFSR 6 h products exhibit good performance in capturing daily variability (Aldersley et al. 2011; Ebisuzaki and Zhang 2011).

Based on a review of the literature (Bonazountas et al. 2005; Thompson and Spies 2009; Braun et al. 2010; Li et al. 2012; Miller and Ager 2013), 10 measurements were selected directly from the CFSR dataset for assessing wildfire susceptibility in Yunnan Province: 6 h average temperature (T_{6h}), 6 h maximum temperature (T_{6h}^{\max}), 6 h minimum temperature (T_{6h}^{\min}), 6 h average surface temperature (ST_{6h}), 6 h precipitation rate ($prate_{6h}$), 24 h precipitation (P_{24h}), 6 h average specific humidity (H_{6h}), 6 h maximum specific humidity (H_{6h}^{\max}), 6 h minimum specific humidity (H_{6h}^{\min}), and 6 h average wind speed (WS_{6h}). The 10 measurements are integrated into 10 days mean/max/min of 6 h and 10 days max and min of 24 h corresponding variables (Table 1) because of the wildfire ignition data are at 10 days temporal scale.

Because the majority of the model inputs are meteorological variables, wildfire point and forest data were rescaled to better incorporate the scale of the meteorological variables (the horizontal resolutions is approximately 38 km). A grid was classified as an ignition grid if at least one wildfire ignition occurred in the grid during the 10 days period, whereas a grid was classified as a nonignition grid if no wildfire ignition occurred. New forest features of a grid due to gridding are characterized by forest coverage ratio and maximum vegetation percentage (Table 1).

The altitude and surface roughness from the CFSR were used as the landform factors in the analysis, while the vegetation classes/subclasses and forest coverage ratio in each cell were used as the vegetation factors. A list of all the variables is shown in Table 1. To illustrate the characteristics of the meteorological and geographical

² <http://www.ncep.noaa.gov/>.

³ <http://apps.ecmwf.int/datasets/data/era40-daily/levtype=sfc/>.

⁴ http://jra.kishou.go.jp/JRA-25/index_en.html.

⁵ <http://gmao.gsfc.nasa.gov/reanalysis/MERRA/>.

⁶ <http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html>.

Table 1 List of variables used in the models

Category	Factor	Meaning of variable	Abbreviation
Temporal factors	Season	Spring, summer, autumn or winter in which wildfire occur	Season
	Month	Jan., Feb., ..., Dec. in which wildfire occur	month
Meteorological factors	Air temperature	10 days mean of 6 h average temperature	$\text{mean}_{10d}\{T_{6h}\}$
		10 days max of 6 h average temperature	$\text{max}_{10d}\{T_{6h}\}$
		10 days min of 6 h average temperature	$\text{min}_{10d}\{T_{6h}\}$
		10 days max of 24 h average temperature	$\text{max}_{10d}\{T_{24h}\}$
		10 days min of 24 h average temperature	$\text{min}_{10d}\{T_{24h}\}$
		10 days mean of 6 h maximum temperature	$\text{mean}_{10d}\{T_{6h}^{\text{max}}\}$
		10 days max of 6 h maximum temperature	$\text{max}_{10d}\{T_{6h}^{\text{max}}\}$
		10 days min of 6 h maximum temperature	$\text{min}_{10d}\{T_{6h}^{\text{max}}\}$
		10 days mean of 6 h minimum temperature	$\text{mean}_{10d}\{T_{6h}^{\text{min}}\}$
		10 days max of 6 h minimum temperature	$\text{max}_{10d}\{T_{6h}^{\text{min}}\}$
		10 days min of 6 h minimum temperature	$\text{min}_{10d}\{T_{6h}^{\text{min}}\}$
		Surface temperature	10 days mean of 6 h average surface temperature
	10 days max of 6 h average surface temperature		$\text{max}_{10d}\{ST_{6h}\}$
	10 days min of 6 h average surface temperature		$\text{min}_{10d}\{ST_{6h}\}$
	10 days max of 24 h average surface temperature		$\text{max}_{10d}\{ST_{24h}\}$
	10 days min of 24 h average surface temperature		$\text{min}_{10d}\{ST_{24h}\}$
	Precipitation rate	10 days mean of 6 h precipitation rate	$\text{mean}_{10d}\{\text{prate}_{6h}\}$
		10 days max of 6 h precipitation rate	$\text{max}_{10d}\{\text{prate}_{6h}\}$
		10 days min of 6 h precipitation rate	$\text{min}_{10d}\{\text{prate}_{6h}\}$
		10 days max of 24 h average precipitation rate	$\text{max}_{10d}\{\text{prate}_{24h}\}$
		10 days min of 24 h average precipitation rate	$\text{min}_{10d}\{\text{prate}_{24h}\}$
	24 h precipitation	10 days mean of 24 h precipitation	$\text{mean}_{10d}\{P_{24h}\}$
		10 days max of 24 h precipitation	$\text{max}_{10d}\{P_{24h}\}$
		10 days min of 24 h precipitation	$\text{min}_{10d}\{P_{24h}\}$
	Specific humidity	10 days mean of 6 h average specific humidity	$\text{mean}_{10d}\{H_{6h}\}$
		10 days max of 6 h average specific humidity	$\text{max}_{10d}\{H_{6h}\}$
		10 days min of 6 h average specific humidity	$\text{min}_{10d}\{H_{6h}\}$
		10 days max of 24 h average specific humidity	$\text{max}_{10d}\{H_{24h}\}$
		10 days min of 24 h average specific humidity	$\text{min}_{10d}\{H_{24h}\}$
		10 days mean of 6 h maximum specific humidity	$\text{mean}_{10d}\{H_{6h}^{\text{max}}\}$
		10 days max of 6 h maximum specific humidity	$\text{max}_{10d}\{H_{6h}^{\text{max}}\}$
		10 days min of 6 h maximum specific humidity	$\text{min}_{10d}\{H_{6h}^{\text{max}}\}$
		10 days mean of 6 h minimum specific humidity	$\text{mean}_{10d}\{H_{6h}^{\text{min}}\}$
		10 days max of 6 h minimum specific humidity	$\text{max}_{10d}\{H_{6h}^{\text{min}}\}$
	Wind speed	10 days mean of 6 h average wind speed	$\text{mean}_{10d}\{WS_{6h}\}$
		10 days max of 6 h average wind speed	$\text{max}_{10d}\{WS_{6h}\}$
10 days min of 6 h average wind speed		$\text{min}_{10d}\{WS_{6h}\}$	
10 days max of 24 h mean of average wind speed		$\text{max}_{10d}\{WS_{24h}\}$	
10 days min of 24 h mean of average wind speed		$\text{min}_{10d}\{WS_{24h}\}$	

Table 1 continued

Category	Factor	Meaning of variable	Abbreviation
Vegetation factors	Vegetation class	Broad-leaved forest, coniferous forest and mixed coniferous broad-leaved forest	Vegetation class
	Vegetation subclass	Subtropical coniferous forest, tropical and subtropical mountain coniferous forest, subtropical broadleaved deciduous forest, subtropical evergreen broad-leaf forest, subtropical monsoon evergreen broad-leaf forest, tropical monsoon forest, hylaea, tropical and subtropical bamboo	Vegetation subclass
	Vegetation	specific tree species of the forest	Vegetation
	Fuel types	Nonflammable, medium flammable, flammable	Fuel types
	Fire spread rate	High speed (>20 m/min), intermediate speed (2–20 m/min), low speed (<2 m/min)	Fire spread rate
	Forest coverage ratio	The ratio between forest covered area and the whole grid	Forest coverage ratio
	Max vegetation percent	Maximum of the percentage of one vegetation covered area and the whole forest covered area	Max vegetation percent
Landform factors	Altitude	Average altitude of the whole grid	Altitude
	Surface roughness	Average surface roughness of the whole grid	Surface roughness

Air temperature and specific humidity are observed 2 m above the ground, while wind speed is observed 10 m above the ground

environment of this study area, the spatial distributions of the main influencing indicators are shown in Fig. 3.

Yunnan Province is divided into 783 grids according to the CFSR meteorological reanalysis data and 419 of the grids cover the forest area, so we have 419 samples from each image. We obtained a 9 years time series of the wildfire ignition data with an interval of 10 days, from 2002 to 2010, with 324 images in total. Hence, there is a total of $419 \times 324 = 135,756$ samples. Among this large sample, there are only 1665 ignition samples, indicating an extreme imbalance between the number of ignition and nonignition samples. To address the imbalance issue, 3330 nonignition samples were randomly selected to form the training samples with the 1665 ignition samples for the ANN, RF, and RF-cost sensitive models. A larger sample of nonignition samples was used because the analysis requires more information to capture the complete distribution of the environmental variables across the entire study area.

3 Methods

We adopted four modeling methods to establish fire ignition models of Yunnan Province: traditional generalized linear models (GLMs)—logistic regression and probit regression, a machine-learning ANN algorithm, and a RF algorithm.

GLMs are extensions of linear regression models that can deal with dependent variables that follow non-normal

distributions (McCullagh and Nelder 1989). For binomial distributions, logistic regression is most frequently used, and probit regression is also often used in natural hazard modeling.

A flow chart demonstrating the estimation of wildfire susceptibility using the various methods is shown in Fig. 4. The flow chart provides information regarding data processing, modeling, and performance evaluation.

3.1 Logistic Regression

Logistic regression is the most fully developed and widely used model to predict qualitative variables, especially binary ones. A logistic regression model assumes that the two-category response variable obeys a binomial distribution, and it specifies a logit link between the dependent and independent variables. The mathematical expression of a logistic regression model is:

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta_1x_1 + \dots + \beta_kx_k$$

where $\pi(x)$ is the expectation of random variable x , especially in the binomial case. $\pi(x)$ is the probability of class “1” (fire ignitions); $1 - \pi(x)$ is the probability of class “0” (nonignitions); and the link function of the model is $g(y) = \log\left(\frac{y}{1-y}\right)$. If we solve the equation, the model also can be expressed as:

$$\pi(x) = \frac{\exp(\alpha + \beta_1x_1 + \dots + \beta_kx_k)}{1 + \exp(\alpha + \beta_1x_1 + \dots + \beta_kx_k)}$$

Fig. 3 Influencing indicators and their spatial distribution. **a** Annual average of temperature, **b** annual average of windspeed, **c** annual average of humidity, **d** annual average of precipitation. **e** altitude, **f** surface roughness, **g** forest coverage ratio, **h** max vegetation percent

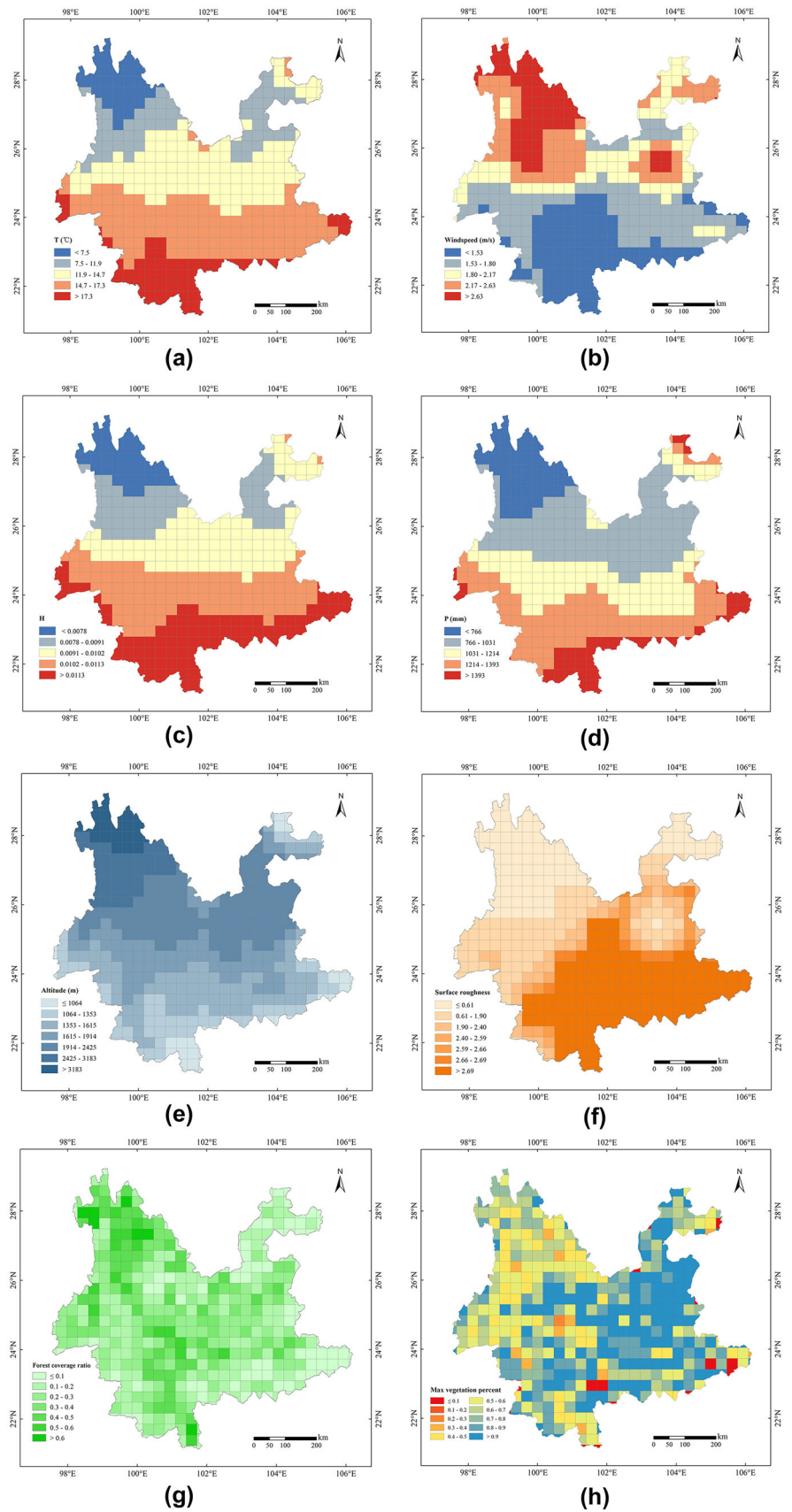
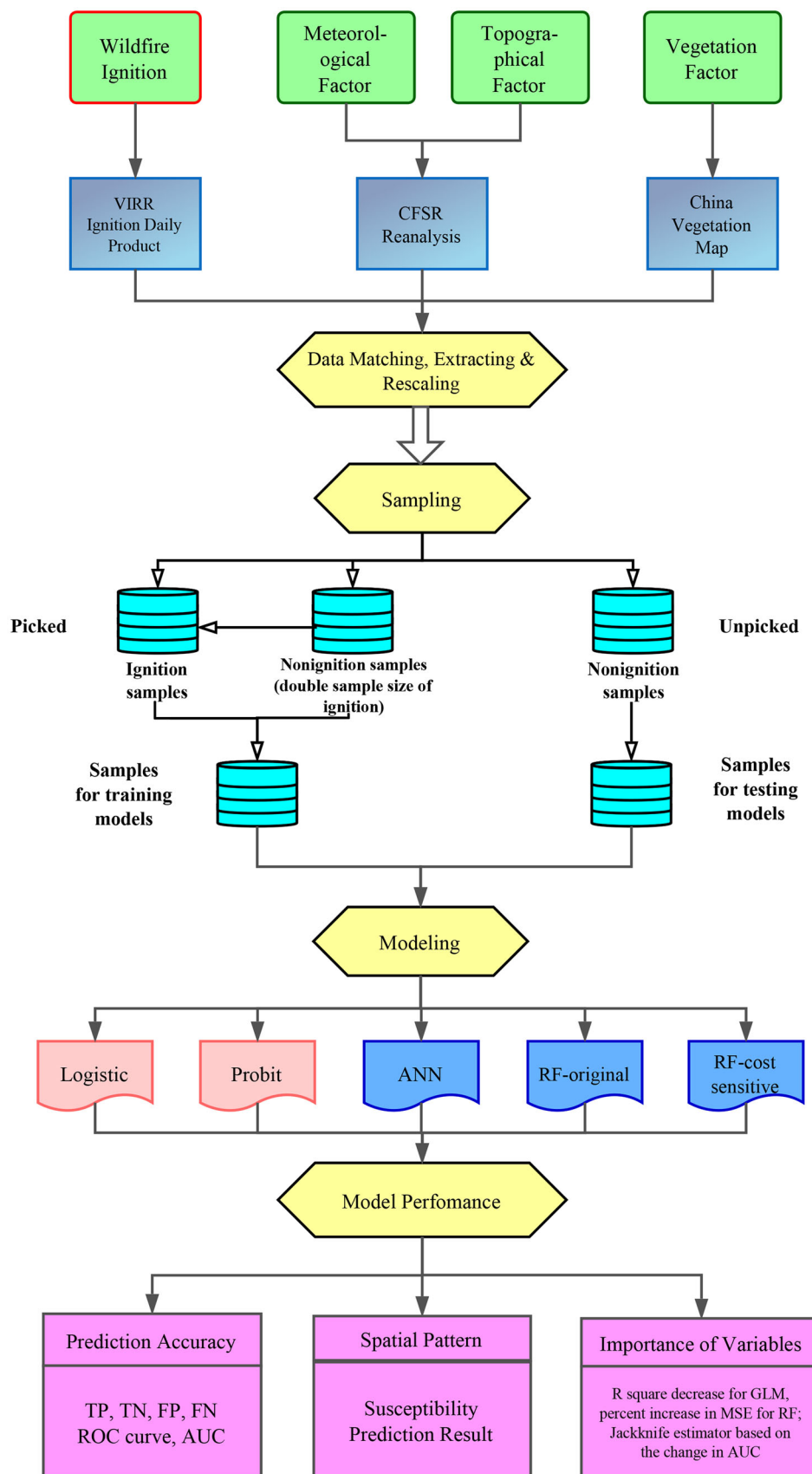


Fig. 4 Flow chart for establishing the wildfire susceptibility models. *Note:* TP is abbreviated for true positive, TN for true negative, FP for false positive, FN for false negative, ROC for receiver operating characteristic, AUC for area under the curve of ROC, GLM for generalized linear models (logistic and probit regression specifically in this article), MSE for mean squared error, RF for random forest algorithm



Considering the strong collinearity of the variables, we developed a logistic regression model using a stepwise procedure, while the best model was selected according to the Akaike information criterion (AIC) (Posada and Buckley 2004; Symonds and Moussalli 2011). Then, we calculated the importance of the variables according to the best model. All of the logistic regression processes were performed using the logistic regression procedure in SAS 9.2 software (SAS Institute 2002–2003).

3.2 Probit Model

A probit model is another generalized regression model that is commonly used to fit binary results. Probit models and logistic regression models both belong to the generalized regression model family. A probit model assumes that the response variable obeys a normal distribution. The mathematical expression of a probit model is:

$$\pi(x) = \Phi(\alpha + \beta_1 x_1 + \dots + \beta_k x_k)$$

where Φ is the cumulative distribution function of the standard normal distribution.

The function curves of probit and logit models are very close to each other so the estimation of the two models can be close. The probit model is more accurate when the dependent variable is normally distributed, and the logit model is more robust when the dependent variable is not normally distributed.

Similar to the logistic regression model, we established the probit model using a stepwise procedure, and we selected the best model based on the AIC. Then, we calculated the importance of the variables according to the best model. All of the probit regression processes were performed using the probit procedure in SAS 9.2 software.

3.3 Artificial Neural Network

Artificial neural networks (ANNs) are a computational model that is loosely analogous to axons in a biological brain. ANNs are widely used in machine learning, computer science, and other research disciplines and are often used as references with which to evaluate the performance of other machine-learning methods. To improve the performance of ANNs, we optimized the structure of the ANN according to the following empirical formula (Wang 2003):

$$l = \log_2 n$$

where l denotes the node number of the hidden layer, and n denotes the node number of the input layer. We established the ANN models using nnet R package (Venables and Ripley 2002).

For testing the model effectiveness, an “out of sample” data subset was created to provide an independent back-

testing view. In the training samples, we randomly selected 1500 out of 1665 fire samples and 330 out of 134,091 non-fire samples, and in the testing samples, the unpicked 165 fire samples and 330 out of 134,091 non-fire samples were used.

3.4 Random Forest

A RF algorithm is basically an ensemble nonlinear classification and regression machine-learning algorithm, which was first proposed by Breiman (2001). The algorithm improves model robustness and prediction accuracy by aggregating classification or regression trees. The algorithm also increases diversity and predictive power by modifying the tree construction method. Each node of the tree is split by the best variable, instead of all of the input variables, from several randomly selected variables.

In terms of model performance, a RF algorithm provides an indicator called an out-of-bag (OOB) error, which is the prediction error of the observations that were not selected by bootstrapping (referred to as OOB data). To evaluate the importance of a variable in the RF algorithm, the difference in the OOB error/mean square error or the Gini index was calculated in each tree when the variable was randomly permuted while all of the other variables remained the same. Then the differences were averaged among all of the trees, and they were used as the measurement indicator of the importance of the variables in the RF algorithm. The difference tells us the extent to which the predictive power of a model is reduced when a given explanatory variable is removed. As detailed in our model, we used the decrease of the OOB error to evaluate the importance of each variable.

Compared with other more frequently used multivariate regression or classification methods, the RF algorithm has several advantages: (1) it is more robust to noise because it randomly selects variables to split at each node; (2) it does not require any assumption regarding the input variables; (3) it allows interactions and nonlinearities among the variables; and (4) it uses the dataset to the utmost. Furthermore, it has a high tolerance of multicollinearity in the input variables, which is very common in wildfire susceptibility assessments.

Because the model input data were extremely unbalanced between ignition and nonignition samples, we also introduced a cost-sensitivity analysis into the RF model (RF-cost sensitive) in addition to the standard RF model (RF-original). The Yunnan wildfire susceptibility models based on RF were established using the R packages randomForest (RF-original) (Liaw and Wiener 2002) and CORElearn (RF-cost sensitive) (Robnik-Sikonja 2004). The randomForest package is based on Breiman and Cutler’s original Fortran code, which was used to implement Breiman’s RF algorithm for classification and regression.

The CORElearn package improves the classification results for imbalanced class data by introducing a cost-sensitive algorithm into the RF algorithm.

3.5 Model Evaluation and Comparison

The performance of the five models was evaluated and compared from three aspects: prediction accuracy, variable importance, and spatial pattern of ignition susceptibility. Here we describe the methods and indicators used for the model evaluation and comparison.

3.5.1 Prediction Accuracy

To compare the precision of the results of the five models, we used the prediction accuracy for each class and the whole model, and the area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Hanley and McNeil 1982) as indicators of model performance. The class prediction accuracy of fire ignitions is also known as the model sensitivity or the true-positive rate (TPR), while the class prediction of nonignitions is also known as the model specificity or the true-negative rate (TNR).

The ROC plot is a graphical representation of the false-positive error rate and the TPR for a binary classification model, and it includes all possible threshold values (Zhou et al. 2009). The x axis of the ROC plot represents the false-positive error rate, which is equal to 1 minus the model specificity, where the model specificity is the class prediction accuracy of nonignitions. The y axis of the ROC plot represents the TPR, which equals the model sensitivity, where the model sensitivity is the class prediction accuracy of fire ignitions. The AUC of the ROC plot is a popular standard metric to assess model classification prediction accuracy because of the ease of interpreting its results, as well as its threshold independence. AUC values range from 0.5 to 1, where 0.5 equals a completely random prediction and 1 means a perfect prediction without misclassification. According to previous research (Bradley 1997; McCune and Grace 2002), an AUC between 0.5 and 0.7 denotes poor model prediction accuracy and performance, an AUC between 0.7 and 0.9 denotes moderate model prediction accuracy and performance, and an AUC greater than 0.9 denotes excellent model prediction accuracy and performance.

3.5.2 Variable Importance Evaluation

We adopted two methods to measure the importance of the variables in the models. The first method was the original variable importance measurement in each model (except ANN, which cannot evaluate variable importance): the R^2 decrease for the GLM and percentage increment in OOB

error for the RF. The percentage increase in the OOB error is the mean of the difference in the OOB error in each tree when the variable is randomly permuted while all of the other variables remain the same. This tells us the extent to which the predictive power of a model is reduced when a given explanatory variable is removed.

The second method was a jackknife estimator of the variable importance based on the change in the AUC using the testing data (Massada et al. 2013). This estimator provides directly comparable results among the models because any binary classification system can be used to calculate ROC curves and to determine the precision of a diagnostic test. The method separately establishes a full model and a partial model (without the variable), and it calculates the AUC using the testing data. The jackknife estimator is the difference between the AUC of the full model and the AUC of the partial model, as the below equation shows:

$$\text{I.O.V.}_j = \frac{1}{n} \sum_{i=1}^n \left(\text{AUC}_{\text{full}}^{(i)} - \text{AUC}_{(j)}^{(i)} \right)$$

where i denotes the test sample that is dropped, and j denotes the variable that the partial model excludes.

The jackknife estimator represents the information provided by a given variable that is not present in the other variables. In addition, we calculated the AUC of the model by excluding one variable at a time, and we compared the AUC values of the models that lacked a single variable, and ranked the variables accordingly.

3.5.3 Spatial Pattern

We calculated every 10 days wildfire ignition susceptibility for each grid using the five models. We used the maximum wildfire ignition susceptibility for the 9 years as the wildfire ignition susceptibility of the grid. Then we compared the wildfire ignition susceptibility spatial patterns predicted by the five models both qualitatively (using graphic images) and quantitatively (by calculating the Spearman correlation coefficient between each pair of maps).

4 Results

Here we present the results from the five models, including two traditional generalized linear models (GLMs)—logistic regression and probit regression, and three machine-learning models—ANN, RF-original, and RF-cost sensitive. The prediction accuracy, variable importance, and spatial pattern of each model result are evaluated and compared, followed by the model sensitivity analysis.

4.1 Prediction Accuracy

Figure 5 shows a four-fold plot illustrating the number of samples and class percentage of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values for the evaluation of the fitting performance of the five models. The proportion of correctly classified samples (TP, TN, TP + TN) is summarized in Table 2.

The GLMs, including the logistic regression and probit regression, performed the worst, with accuracies of only 77.60 and 78.20%, respectively, for wildfire ignition and 66.24 and 65.99%, respectively, for all samples. The ANN performed well when predicting wildfire ignitions, with an accuracy of 83.78%, but it performed poorly when predicting nonignitions, with an accuracy of only 78.44%. The RF-original exhibited the most balanced prediction accuracy, with an accuracy of 84.26% for ignitions and 88.35% for nonignitions. The RF-cost sensitive model performed the best, with an accuracy of 88.47% for all of the samples and 94.23% accuracy for wildfire ignition prediction. Compared with traditional models, such as the logistic regression model, the probit model, and the ANN, the RF-cost sensitive model increased the total accuracy by 22.23, 22.48, and 9.56%, respectively, and by 16.63, 16.03, and 10.45%, respectively, for the wildfire ignition prediction.

All five models exhibited excellent prediction performances. We drew ROC curves of the five models, as shown in Fig. 6. The ROC curves clearly show that the RF-cost sensitive model had the best performance, with an AUC of 0.9848, followed by the RF-original model, the ANN, the probit regression model, and the logistic regression model, with AUCs of 0.9651, 0.9346, 0.9297, and 0.9296, respectively.

4.2 Importance of the Variables

The variable importance rank of each model is listed in Table 3. The results show that forest coverage ratio, month, season, surface roughness, 10 days minimum of the 6 h maximum specific humidity ($\min_{10d}\{H_{6h}^{\max}\}$), and 10 days maxima of the 6 h average and maximum temperatures ($\max_{10d}\{T_{6h}\}$ and $\max_{10d}\{T_{6h}^{\max}\}$, respectively) contributed most to the susceptibility to wildfire ignition. Moreover, the seven most important variables appeared in all four of the optimal models, indicating that they are dominant influencing variables of wildfire ignition under different modeling methods.

The most important variable, forest coverage ratio, which is the ratio between the forest cover area and the area of the whole grid, was the most important variable in the GLMs, and it was very important in the RFs. The forest coverage ratio accounted for 0.0454 and 0.0449 of the R^2

values in the logistic regression and probit models, respectively, and it accounted for approximately 6.9 and 6.8%, respectively, of the R^2 values of the full models. The forest coverage ratio also accounted for 3.24 and 2.69% of the OOB error decrease in the RF-original and RF-cost sensitive models, respectively.

The second important variable, month, was most important in the RFs, and it was very important in the GLMs. The third most important variable, season, ranked second or third in all four models. Together, month and season accounted for 0.0223 and 0.0253 of the R^2 values in the logistic regression and probit models, which accounted for approximately 3.4 and 3.8%, respectively, of the R^2 values of the full model. They also collectively accounted for 12.67 and 12.10% of the OOB error decrease in the RF-original and RF-cost sensitive models, respectively. The high importance of month and season very likely results from the distinct monthly and seasonal patterns of wildfire ignitions (Fig. 2b) in which winter and spring accounted for an overwhelming majority of the wildfire ignitions. Detailed by month, we observed a gradual increase in wildfire ignitions from December to March (peak), and a gradual decrease from March to May. The pattern is mainly the result of the distinct dry and wet seasons in Yunnan Province.

The rainy season in Yunnan usually starts from the mid-May and lasts until the end of October, and the dry season starts in early November and lasts until around 20 May. At the beginning of the dry season, although the precipitation declines substantially, there is sufficient water in plants and soil because of the accumulation of water during the rainy season and the low evaporation rate that results from low temperatures. Therefore there were few wildfire ignitions in November because the forest was reasonably humid. As the dry season continues, the plants and soil continue to lose water, and the forest dries out until wildfires break out in December and increase in number in January. By the following spring, most of Yunnan is under a dry and warm tropical continental air mass. The steady air mass leads to warmer temperatures, more sunny days, and greater evaporation, which results in continuously accelerating water loss. The wildfire ignition risk increased significantly in February and peaked in March. After April, the rainfall intensity and frequency increase gradually in Yunnan as the wet season approaches. Hence, the wildfire ignition risk decreased significantly until June, when there were only a few wildfire ignitions in Yunnan. In July through September (the middle of the wet season), no wildfire ignitions were observed in Yunnan Province.

In addition, temperature and moisture were vital drivers of wildfire ignitions. The 10 days minimum of the 6 h maximum specific humidity ($\min_{10d}\{H_{6h}^{\max}\}$) in each grid

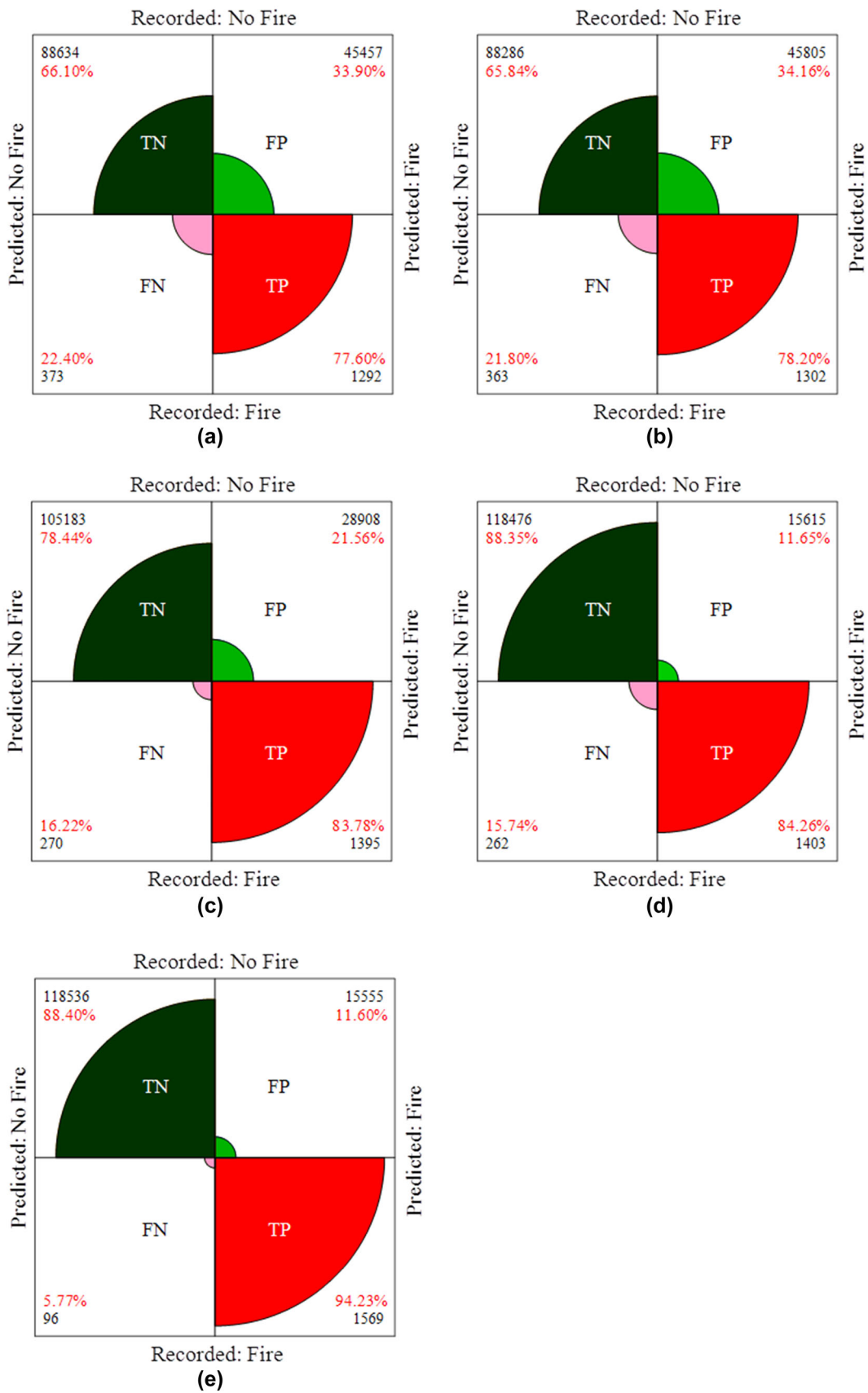


Fig. 5 Fitting performance of the five models. Number and class percentage of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) values for the evaluation of the fitting performance of the five models in the fourfold figures: **a** logistic regression, **b** probit regression, **c** ANN, **d** RF-original, **e** RF-cost sensitive

Table 2 Class prediction accuracy of each model

	Ignition (%)	Nonignition (%)	Total accuracy (%)
Logistic	77.60	66.10	66.24
Probit	78.20	65.84	65.99
ANN	83.78	78.44	78.51
RF-original	84.26	88.35	88.30
RF-cost sensitive	94.23	88.40	88.47

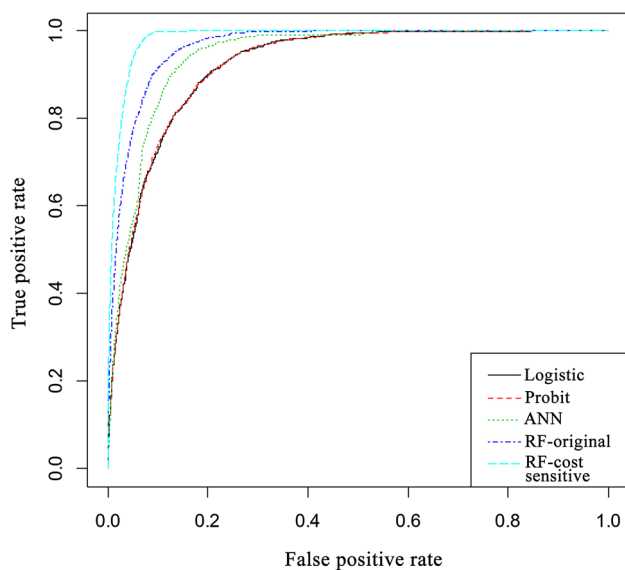


Fig. 6 ROC curves of the five models

ranked fifth in terms of importance to wildfire ignition susceptibility; the 10 days maxima of the 6 h average and maximum temperatures ($\max_{10d}\{T_{6h}\}$ and $\max_{10d}\{T_{6h}^{\max}\}$) of the grid ranked sixth and seventh, respectively.

4.3 Spatial Pattern

Figure 7 shows the spatial patterns of wildfire ignition susceptibility estimated by the five models. Southern Yunnan, Wenshan and Baoshan, and some areas south of the Hengduan Mountain exhibited the highest wildfire ignition susceptibilities. All of the models, except the ANN, predicted the two high ignition susceptibility areas,

while the ANN only predicted a small part of southern Yunnan to have high susceptibility.

We observed a similar result from the Spearman correlation coefficient of each pair of the five models (Table 4). The two GLMs (the logistic regression and probit models) exhibited highly correlated results. The GLMs and RFs produced similar wildfire ignition susceptibility maps with strong correlations (Spearman correlation coefficients >0.75), and the ANN produced a different map that was only moderately correlated with the rest of the models. This corresponds to the comparison of the susceptibility maps shown in Fig. 7.

The map of wildfire ignition susceptibility can be further used to generate an ignition probability map by using the Monte Carlo approach (Metropolis 1987; Ye et al. 2017). To illustrate this, we used the annual ignition data from 1990 to 2015 of Yunnan Province to fit a distribution and randomly generated the number of annual ignitions for 1000 years. The specific location point of each wildfire event was generated based on ignition susceptibility maps derived from Fig. 7e by using the acceptance-rejection method (Casella et al. 2004). Figure 7f shows some spatial similarity to the ignition susceptibility map but it provides information about the probability of ignition.

4.4 Model Sensitivity

The optimization of the models may also influence their prediction accuracy and model performance. Apart from the GLMs with fixed optimal method, the optimization of the ANN and RF algorithm needs more discussion. In the ANN model, the node number of the input layer was $n = 50$; according to the empirical formula (Wang 2003), $l = \log_2 n$, the optimized node number of the hidden layer was close to $l = 6$. Therefore we simulated $l = 5-9$, and for each l , we built 100 ANN models. The highest accuracies occurred when $l = 5$ or 9, but considering the prediction of ignition, $l = 6$ was the best model, although it had a rather high failure ratio (Table 5). But the main shortcoming of the ANN is its lack of robustness. The failure ratio in Table 5 represents the ratio of the ANN models that predicted that all of the samples belonged to the nonignition class, almost more than half of the models cannot predict at all.

We simulated the RF-original models with different number of trees (Table 6). The table shows that the total accuracy of all of the models was greater than 80%, and when the number of trees was greater than 20, the total accuracy was greater than 85%. The maximum total accuracy was 86.85% for 80 trees, which is more than the 85.57% accuracy for 20 trees. The increased precision is not worth it, given the increased computation complexity.

Table 3 List of the seven most important variables in the four models under different measurement methods

Variable	Logistic		Probit		RF-original		RF-cost sensitive		Rank (average)
	Rsq (rank)	AUC (rank)	Rsq (rank)	AUC (rank)	OOB (rank)	AUC (rank)	OOB (rank)	AUC (rank)	
Forest coverage ratio	0.0454 (1)	0.016 (1)	0.0449 (1)	0.0161 (1)	0.0324 (4)	0.0009 (6)	0.0269 (6)	0.0010 (2)	1 (2.75)
Month	0.0054 (4)	0.0019 (5)	0.0061 (6)	0.0022 (6)	0.0807 (1)	0.0012 (3)	0.0723 (1)	0.0004 (4)	2 (3.75)
Season	0.0169 (3)	0.0044 (3)	0.0192 (3)	0.0051 (3)	0.0460 (2)	-0.0001 (12)	0.0487 (2)	-0.0003 (14)	3 (5.25)
Surface roughness	0.0027 (7)	0.0008 (11)	0.0045 (9)	0.0013 (9)	0.0258 (6)	0.0018 (1)	0.0306 (5)	0.0007 (3)	4 (6.38)
$\min_{10d}\{H_{6h}^{max}\}$	0.0016 (12)	0.0004 (15)	0.0069 (5)	0.0023 (5)	0.0235 (11)	0.0011 (5)	0.0322 (4)	-0.0003 (13)	5 (8.75)
$\max_{10d}\{T_{6h}\}$	0.0025 (9)	0.0008 (9)	0.0053 (8)	0.0020 (8)	0.0246 (8)	-0.0003 (15)	0.0257 (8)	-0.0002 (12)	6 (9.63)
$\max_{10d}\{T_{6h}^{max}\}$	0.0031 (6)	0.001 (7)	0.0007 (17)	0.0003 (17)	0.0253 (7)	0.00004 (8)	0.0239 (11)	0.0001 (7)	7 (10.00)

Numbers represent the decrease of R^2 /AUC or the increment of OOB error and rank of each variable

In conclusion, the OOB error for the model prediction precision was not sensitive to the number of trees.

We trained the RF-cost sensitive model using different ratios of ignition and nonignition samples to find the best sample ratio, the results of which are shown in Table 7. As the number of nonignition samples increased, the accuracy of ignition prediction decreased, while the accuracy of nonignition and total predictions increased. Considering the balance between the prediction accuracy of the ignition and nonignition samples, 1:2 is the best ratio for the ignition and nonignition samples.

5 Discussion

The purpose of evaluation and comparison of the models is to find appropriate models for better prediction accuracy and better understanding of regional wildfire ignition susceptibility. In this study, we show that susceptibility is highly influenced by meteorological variables such as humidity and precipitation. Regional drought in Yunnan has a connection with wildfire ignition susceptibility and we discuss this susceptibility in greater detail, as well as explore forest management practice using the analyzed ignition susceptibility.

5.1 Wildfire and Drought

Yunnan’s yearly variation in wildfire (Fig. 2b) is considerable, because there is a very strong connection between the number of wildfire ignitions and drought. Severe droughts in winter and spring, early summer, or midsummer occurred continuously from 2003 to 2007. The drought in the spring and summer of 2005 and the spring drought in

2006 were the most severe droughts in the last 50 and 20 years, respectively. Frequent droughts yielded a continuous increase in the wildfire ignition number from 2003 to 2006. We speculate that there are three reasons for the unexpected decrease of wildfire ignitions in 2007: (1) the winter and spring droughts were obviously less severe than those in 2005 and 2006; (2) the spring drought in 2006 caused two wildfires and consumed a great deal of the combustible material in the forests, while the drought suppressed the germination and growth of forest plants and further reduced the amount of combustible material in the forests; and (3) the reduction of combustible material that resulted from the abundant wildfires in 2003–2005 cannot be ignored. In contrast, the severe freezing rain and snow events in early 2008 greatly reduced the number of winter wildfires, at the same time that winter moisture reduced the number of spring wildfires to the same level observed in 2003. Yet these precipitation events also produced unprecedented amounts of combustible material, which led to a large increase in the number of wildfires in subsequent years. This wildfire increase was enhanced by the spring and summer drought in 2009 and the severe once-in-a-century drought in 2010.

Studies of Yunnan droughts have found that in the twenty-first century droughts between September and December have been more severe than those occurring before 2000. Droughts between January and March have remained at the same level since 2000, but have been more severe than the droughts of the 1990s (Zhang et al. 2013). Both these trends led to more frequent, intense, and continuous winter to spring droughts, which aggravate wildfires in Yunnan. There is evidence that more megadroughts occurred historically in those parts of Asia (including Yunnan) that are subject to the influence of the tropical

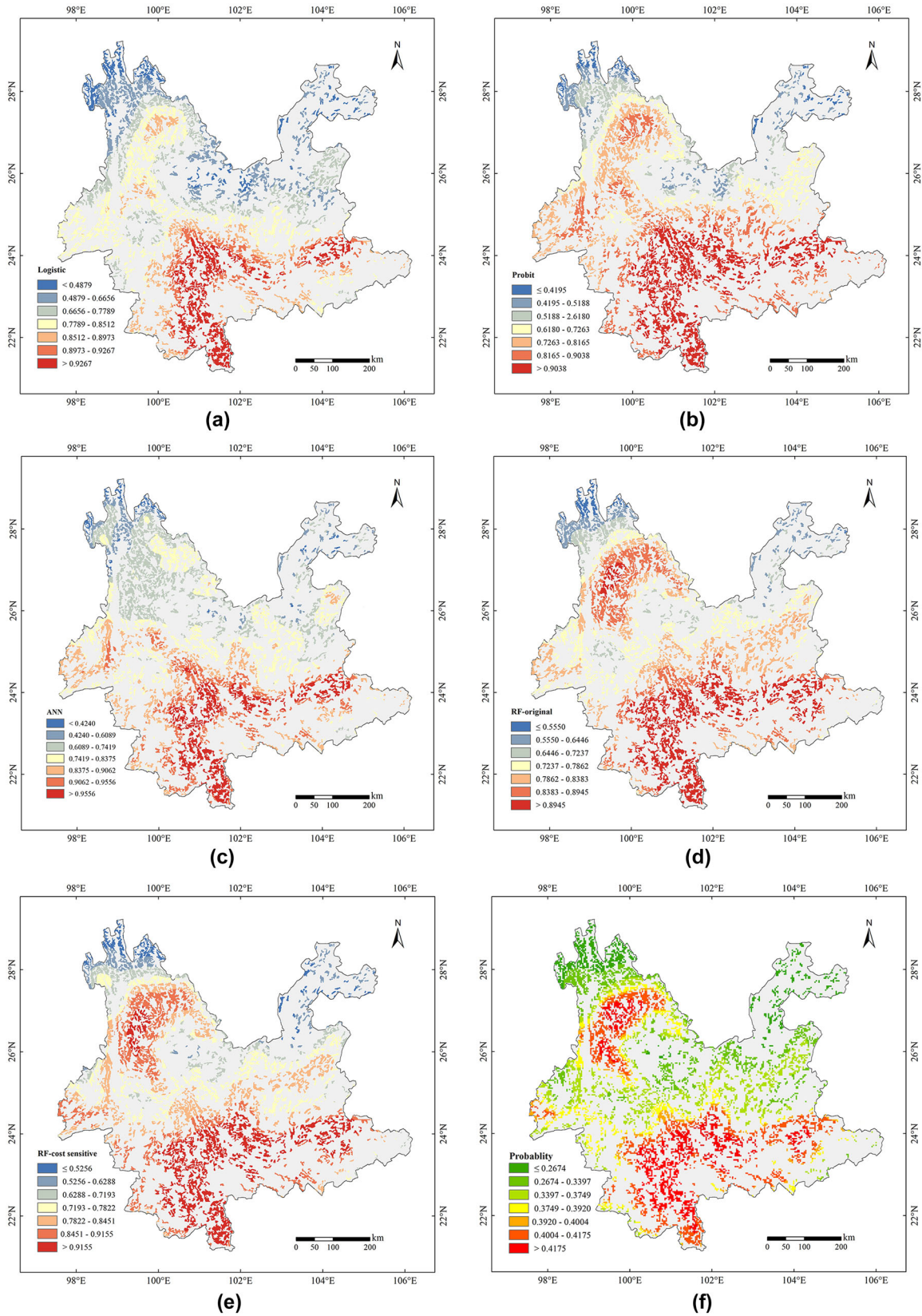


Fig. 7 Spatial pattern of susceptibility for each model: **a** logistic regression, **b** probit regression, **c** the ANN, **d** RF-original, and **e** RF-cost sensitive, **f** depicts ignition probability

Table 4 Spearman correlation coefficient of the wildfire ignition susceptibility spatial pattern

	Probit	ANN	RF-original	RF-cost sensitive
Logistic	0.98435	0.39593	0.75690	0.80649
Probit		0.47432	0.77369	0.81266
ANN			0.49771	0.44432
RF-original				0.93499

Table 5 Prediction precision of the ANN models

	Ignition	Nonignition	Total accuracy	Failure ratio
$l = 5$	0.831692	0.793520	0.793988	0.70
$l = 6$	0.896734	0.767605	0.769189	0.68
$l = 7$	0.855210	0.789178	0.789988	0.60
$l = 8$	0.866230	0.779119	0.780187	0.56
$l = 9$	0.864068	0.793745	0.794607	0.48

Failure ratio means the ratio of the ANN models that predict all samples as belonging to the nonignition class. Ignition represents the average prediction precision of the fire samples, while nonignition represents the average prediction precision of the nonignition samples; total accuracy stands for the average predict precision of all samples. All of the averages were calculated without failure models

Table 6 Prediction precision of the RF-original models

Number of trees	Total accuracy
10	0.8262262
20	0.8520521
30	0.8556557
40	0.8628629
50	0.8592593
60	0.8634635
70	0.8632633
80	0.8684685
90	0.8672673
100	0.8662663

Table 7 Prediction precision of different sample ratios using the RF-cost sensitive model

	Ignition	Nonignition	Total accuracy
No cost-matrix	0.8426426	0.8835492	0.8830
1:1	0.9723724	0.8043567	0.8064174
1:2	0.9423423	0.8839743	0.8846902
1:3	0.8960961	0.9217770	0.9214620
1:4	0.8606607	0.9488034	0.9477224
1:10	0.6282282	0.9883885	0.9839712

monsoon. This trend was particularly pronounced during the Little Ice Age (1450–1850) compared with the Medieval Climate Anomaly between 950 and 1250 (IPCC 2013). Additionally, since 2000, Yunnan already has experienced two megadroughts, one in 2005 and one from 2009 to 2011. All of this reveals an extremely high risk of megadroughts in Yunnan. Considering the strong link between drought and wildfire, megadroughts can increase the frequency and severity of wildfires, as well as expand the area affected by wildfires.

5.2 Wildfire Susceptibility and Forest Management

The spatial patterns of the wildfire susceptibility estimated by the five models (Fig. 7) indicate that the high susceptibility areas in Yunnan are southern Yunnan, Wenshan and Baoshan, and some areas south of the Hengduan Mountains. The high susceptibility of southern Yunnan and Wenshan and Baoshan can be explained by their higher forest fuel loads than other areas. These areas have the best water and heat condition and the highest vegetation density, which leads to faster accumulation of combustible material. The Hengduan Mountains play an important role in the spatial pattern of wildfire susceptibility in the region. Yunnan is controlled by warm and dry westerly winds and a northerly air current in the dry season. When the westerly wind passes over the Hengduan Mountains, the air current descends, leading to an increase in air temperature and a decrease in humidity. At the same time, the mountain ranges block the entry of cold air from the north and warm and moist air from the south, which results in very limited rainfall. Therefore, the Hengduan Mountains intensify wildfire susceptibility in Yunnan, especially in the southern area of the Hengduan Mountains.

The wildfire susceptibility maps generated in this study provide objective and clear guidance for wildfire management in Yunnan from the aspect of spatial variation. Special attention should be paid to the region's high susceptibility areas. The most important variables include temporal, vegetation, and meteorological factors that are directly linked to the wildfire susceptibility. These variables must be correctly understood and perceived by regional and local wildfire managers in order to mitigate the area's serious wildfire hazard.

6 Conclusions

This study has confirmed that of the five models investigated, the RF-cost sensitive analysis was the best method for predicting wildfire ignition susceptibility. The RF-cost sensitive analysis had the highest accuracy (88.47%) for all of the samples, and 94.23% accuracy for wildfire ignition

prediction in Yunnan. Compared with widely used GLM models (logistic regression and probit regression models) and the ANN, the RF-original model increased total accuracy by 22.23, 22.48, and 9.56%, respectively, and the wildfire ignition prediction by 16.63, 16.03, and 10.45%, respectively. Wildfire susceptibility can be assessed using various models, which range from conventional regressions to more recently developed machine-learning models. Careful processing of data samples is needed, however, to resolve issues of data imbalance and to avoid potentially misleading results due to the overwhelmingly large number of nonignition samples. The tradeoff between overall accuracy and ignition prediction also needs special attention. High sensitivity (the TPR) should be obtained for good ignition prediction, and the specificity (the TNR) and accuracy factors should also be given consideration. Although the performance of machine-learning methods (the ANN and RF models) investigated in this study was better than that of the logistic and probit regressions, the numbers of layers in the ANN and trees in the RF should be further tested to achieve optimized results.

The importance of variables in the RF models indicates that the factors mainly influencing Yunnan wildfire occurrence are forest coverage ratio, month/season, surface roughness, 10 days minimum of the 6 h maximum specific humidity, as well as the 10 days maxima of the 6 h average and maximum temperatures. The most susceptible areas were located in southern Yunnan and some areas south of the Hengduan Mountains. Under a future global warming scenario, wildfire susceptibility in Yunnan could further increase, particularly in terms of frequency and duration. Grow of this potential hazard threat is a result of the increasing severity of continuous, large-scale winter to spring droughts in the region. The identified dominant influencing factors help us to better understand wildfire occurrence, and the developed susceptibility map provides guided spatial information for regional wildfire risk management.

Acknowledgement This work has been supported by the international partnership program of Chinese Academy of Sciences (Grant # 131551KY5B20160002) and the National Natural Science Foundation of China (Grants # 41671503 and 41621061).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aldersley, A., S.J. Murray, and S.E. Cornell. 2011. Global and regional analysis of climate and human drivers of wildfire. *Science of the Total Environment* 409(18): 3472–3481.
- Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7): 1145–1159.
- Braun, W.J., B.L. Jones, J.S.W. Lee, D.G. Woolford, and B.M. Wotton. 2010. Forest fire risk assessment: An illustrative example from Ontario, Canada. *Journal of Probability and Statistics* 2010: Article No. 823018.
- Bonazountas, M., D. Kallidromitou, P.A. Kassomenos, and N. Passas. 2005. Forest fire risk analysis. *Human and Ecological Risk Assessment: An International Journal* 11(3): 617–626.
- Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2): 123–140.
- Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32.
- Casella, G., C.P. Robert, and M.T. Wells. 2004. Generalized accept–reject sampling schemes. *Lecture Notes-Monograph Series* 45: 342–347.
- Catani, F., D. Lagomarsino, S. Segoni, and V. Tofani. 2013. Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues. *Natural Hazards and Earth System Science* 13(11): 2815–2831.
- Chen, F., X.-D. Lin, S.-K. Niu, S. Wang, and D. Li. 2012. Influence of climate change on forest fire in Yunnan Province, southwestern China. *Journal of Beijing Forestry University* 34(6): 7–15 (**in Chinese**).
- Chen, F., Z.F. Fan, S.K. Niu, and J.M. Zheng. 2014. The influence of precipitation and consecutive dry days on burned areas in Yunnan Province, southwestern China. *Advances in Meteorology*. doi:10.1155/2014/748923.
- Del Río, S., V. López, J.M. Benítez, and F. Herrera. 2014. On the use of MapReduce for imbalanced big data using random forest. *Information Sciences* 285: 112–137.
- Domingos, P. 1999. Metacost: A general method for making classifiers cost-sensitive. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, USA, 15–18 August 1999. 155–164. New York: ACM.
- Ebisuzaki, W., and L. Zhang. 2011. Assessing the performance of the CFSR by an ensemble of analyses. *Climate Dynamics* 37(11–12): 2541–2550.
- Elkan, C. 2001. The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Seattle, Washington, USA, 4–10 August 2001. ACM Digital Library. San Francisco, CA: Morgan Kaufmann Publishers.
- Fang, K.N.B., and W.U. Jian-Bina. 2011. A review of technologies on random forests. *Statistics & Information Forum* 2011(3): 33–39.
- Finney, M.A. 2005. The challenge of quantitative risk analysis for wildland fire. *Forest Ecology and Management* 211(1): 97–108.
- Hagan, M.T., H.B. Demuth, and M.H. Beale. 1996. *Neural network design*. Boston: PWS Publishing.
- Hanley, J.A., and B.J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1): 29–36.
- Hsu, K.L., H.V. Gupta, and S. Sorooshian. 1995. Artificial neural network modeling of the rainfall–runoff process. *Water Resources Research* 31(10): 2517–2530.

- Huang, B.W. 1989. The comprehensive regionalization compendium of Chinese nature. *Collection of Geographical Publications* 21: 10–20 (in Chinese).
- IPCC (Intergovernmental Panel on Climate Change). 2013. *Climate change 2013. The physical science basis*. Contribution of Working Group I to the fifth assessment report of the Intergovernmental Panel on Climate Change, ed. T.F. Stocker, D. Qin, G.K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley. Cambridge, UK: Cambridge University Press.
- Li, X.W., G.B. Fu, M.J.B. Zeppel, X.B. Yu, G. Zhao, D. Eamus, and Q. Yu. 2012. Probability models of fire risk based on forest fire indices in contrasting climates over China. *Journal of Resources and Ecology* 3(2): 105–117.
- Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. *R News* 2(3): 18–22.
- Massada, A.B., A.D. Syphard, S.I. Stewart, and V.C. Radeloff. 2013. Wildfire ignition-distribution modelling: A comparative study in the Huron-Manistee National Forest, Michigan, USA. *International Journal of Wildland Fire* 22(2): 174–183.
- McCullagh, P., and J.A. Nelder. 1989. *Generalized linear models*. Vol. 37. London: CRC Press.
- McCune, B., and J.B. Grace. 2002. *Analysis of ecological communities*. Glenden Beach, OR: MjM Software Design.
- Metropolis, N. 1987. The beginning of the monte-carlo method. *Los Alamos Science* No. 15(Special Issue): 125–130.
- Miller, C., and A.A. Ager. 2013. A review of recent advances in risk analysis for wildfire management. *International Journal of Wildland Fire* 22(1): 1–14.
- Oliveira, S., F. Oehler, J. San-Miguel-Ayanz, A. Camia, and J.M.C. Pereira. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *Forest Ecology and Management* 275: 117–129.
- Peng, G.F., Y. Liu, and Y.P. Zhang. 2009. Research on characteristics of drought and climatic trend in Yunnan Province. *Journal of Catastrophology* 24(4): 40–44 (in Chinese).
- Posada, D., and T.R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* 53(5): 793–808.
- Robnik-Sikonja, M. Improving random forests. Paper presented at the ECML 2004, Pisa, Italy, 20–24 September 2004.
- Rodrigues, M., and J. de la Riva. 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software* 57: 192–201.
- Saha, S., S. Moorthi, H.-L. Pan, X.R. Wu, J. Wang, S. Nadiga, P. Tripp, R. Kistler, et al. 2010. NCEP climate forecast system reanalysis (CFSR) selected hourly time-series products, January 1979 to December 2010. Boulder, CO: Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory.
- SAS Institute. 2002–2003. *SAS system version 9.2 for Windows*. Cary, NC: SAS Institute.
- State Forestry Administration. 1990–2015. *China forestry statistical yearbook*. Beijing: China Forestry Press (in Chinese).
- Sun, Y.M., M.S. Kamel, A.K.C. Wong, and Y. Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40(12): 3358–3378.
- Symonds, M.R.E., and A. Moussalli. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology* 65(1): 13–21.
- Thompson, J.R., and T.A. Spies. 2009. Vegetation and weather explain variation in crown damage within a large mixed-severity wildfire. *Forest Ecology and Management* 258(7): 1684–1694.
- Venables, W.N., and B.D. Ripley. 2002. *Modern applied statistics with S-PLUS*. New York: Springer.
- Wang, S.C. 2003. Artificial neural network. In *Interdisciplinary Computing in Java Programming*, 81–100. Boston, MA: Springer.
- Weinstein, D., and P. Woodbury. 2010. Review of methods for developing probabilistic risk assessments. Part 1: Modeling fire. *Advances in threat assessment and their application to forest and rangeland management* 2: 285–302.
- Ye, T., and Y. Wang, Z.X. Guo, and Y.J. Li. 2017. Factor contribution to fire occurrence, size, and burn probability in a subtropical coniferous forest in East China. *PLoS ONE* 12(2): e0172110.
- Zhang, M.J., J.Y. He, B.L. Wang, S.J. Wang, S.S. Li, W.L. Liu, and X.N. Ma. 2013. Extreme drought changes in Southwest China from 1960 to 2009. *Journal of Geographical Sciences* 23(1): 3–16.
- Zhang, X.S. (ed.). 2007. *Vegetation map of the People's Republic of China (1:1000000)*. Beijing: Geology Press (in Chinese).
- Zhou, X.H., D.K. McClish, and N.A. Obuchowski. 2009. *Statistical methods in diagnostic medicine*. Vol. 569. New York: Wiley & Sons.