Evaluating lightning-caused fire occurrence using spatial generalized additive models: a case study in Central Spain

José Ramón Rodríguez-Pérez[1], Celestino Ordóñez[2], Javier Roca-Pardiñas[3], Daniel Vecín-Arias[1], Fernando Castedo-Dorado[1]*

[1] Universidad de León, GEOINCA Research Group, Ponferrada (León), SPAIN

[2] Universidad de Oviedo, Department of Mining Exploitation and Prospecting, Polytechnic School of Mieres, Mieres (Asturias), SPAIN

[3] Universidad de Vigo, Department of Statistics and Operations Research, SIDOR Research Group, Vigo (Pontevedra), SPAIN.

*Address correspondence to Fernando Castedo-Dorado, GEOINCA Research Group, Universidad de León, Campus de Ponferrada, Avda. de Astorga SN; tel: +34-987442028; fax: +34-987442000; fcasd@unileon.es

Abstract

It is widely accepted that the relationship between lightning wildfire occurrence and its influencing factors vary depending on the spatial scale of analysis, making the development of models at the regional scale advisable. In this study we analyse the effects of different biophysical variables and lightning characteristics on lightning-caused forest wildfires in Castilla y León region (Central Spain). The presence/absence of at least one lightning-caused fire in any 4 × 4-km grid cell was used as a dependent variable and vegetation type and structure, terrain, climate and lightning characteristics were used as possible covariates. Five prediction methods were compared: a generalized linear model (GLM), a random forest model (RFM), a generalized additive model (GAM), a generalized additive model that includes a spatial trend function (GAMs) and a spatial autoregressive model (AUREG).

A GAMs with just one covariate, apart from longitude and latitude for each observation included as a combined effect, was considered the most appropriate model in terms of both predictive ability and simplicity. According to our results, the probability of a forest being affected by a lightning-caused fire is positively and non-linearly associated with the percentage of coniferous woodlands in the landscape, suggesting that occurrence is more closely associated with vegetation type than with topography, climate or lightning characteristics.

The selected GAMs is intended to inform the Regional Government of Castilla y León (the fire and fuel agency in the region) regarding identification of areas at greatest risk so it can design long-term forest fuel and fire management strategies.

Keywords: lightning-caused fires; spatial generalized additive models; lightning fire occurrence; spatial effect

200-character summary

A GAMs that included the percentage of coniferous woodlands as the covariate best explained and predicted long-term lightning wildfire occurrence in Central Spain.

1. INTRODUCTION

Wildfires have significant effects on life, property and the environment worldwide. Highly damaging wildfire events have caused a major loss of human lives and of forested ecosystems in Mediterranean Europe in recent decades (Molina-Terrén et al., 2019; Moreira et al., 2011). A critical component of wildfire risk is a better understanding of both natural (lightning) and human fire ignition sources. Fire risk estimates are therefore crucial to pre-empt and reduce the negative impact of wildfires. This requires the development of occurrence likelihood models in order to understand and estimate risk.

Wildfires caused by lightning (lightning-caused wildfires) are a particularly important problem in boreal forests, where they represent over 70% of the total burned area (Flanningan & Wotton, 1991; McGuiney, Shulski, & Wendler, 2005). In recent years in the Mediterranean basin, although not the main cause, lightning has become a more frequent cause of wildfires (Vázquez & Moreno, 1998). In the 2001-2010 period, for instance, around 25% of wildfires that burned more than 3000 ha were caused by lightning (MAGRAMA, 2012). Nevertheless, due to the greater significance of people-caused wildfires in Spain and other Mediterranean countries, little attention has been paid to the modelling of lightning-caused wildfires.

Wildfires are a major disruptive agent in the natural environment of Castilla y León in central Spain, among the largest self-governed regions in Europe, with a total of 94,213 km^2 (nearly 20% of Spanish territory), of which some 50,000 km^2 are forested. The consequences of wildfires in terms of risks to the population and economy have been exacerbated in this region in recent decades by important socioeconomic transformations, including land abandonment and urban pressures on forested areas (the wildland-urban interface), resulting in an increase in wildfire spread and severity and in vulnerability. On average, 8% of the wildfires that occur in this region are due to lightning strikes during thunderstorms, although in some areas they have comprised more than 50% of the total wildfires in a year (Martínez, Martínez-Vega, & Martín,

2004). Additionally, in some years lightning-caused wildfires account for more than 20% of the total burned area in the region (MAGRAMA, 2012).

It is widely accepted that the variables that potentially influence the spatial distribution of lightning-caused fires are related to vegetation type and structure, terrain, weather and lightning characteristics (Krawchuk, Cumming, Flannigan, & Wein, 2006). In comparison to human-induced fires in central Spain, lightning-induced fires are less frequent and fire rotation periods are therefore usually high; lightning fires also usually affect woodlands more and start under different meteorological conditions (Vázquez & Moreno, 1998).

Lightning-caused fire occurrence models can be classified as short-term or long-term on the basis of temporal resolution (Chuvieco, Salas, Carvacho, & Rodríguez-Silva, 1999). Short-term wildfire occurrence models are largely dependent on weather conditions, which directly affect fuel moisture and lightning activity. Long-term wildfire occurrence models refer to more permanent factors associated with fire ignition such as topography, vegetation composition and structure, climate and lightning patterns (San Miguel-Ayanz et al., 2003). Identifying areas of high lightning-caused fire risk in short time scales (e.g., daily) can assist fire management agencies in shifting resources between localities to ensure firefighting needs are met (Chow & Regan, 2011). Long-term assessment focuses on investigating the structural factors that affect the fire proneness of an area and so helps define prevention strategies, e.g., identifying areas where fire detection efforts or fuel treatments need to be intensified or determining the long-term allocation of firefighting resources (Oliveira, Oehler, San Miguel-Ayanz, Camia, & Pereira, 2012; San Miguel-Ayanz et al., 2003).

Prevention activities in Castilla y León are, inter alia, focused on (i) landscape vigilance from lookout towers with the aim of minimizing time to detection of incipient wildfires, (ii) management of surface fuels through shrub mastication and prescribed burning, and (iii) management of crown fuels through thinning and pruning. Due to financial constraints, the

latter two fuel management approaches are not implemented on an extensive scale (area-wide) but are strategically focused on certain stands or strips. Long-term identification of areas prone to lightning ignition is therefore useful for operational purposes, as it ensures more effectively focused prevention actions. Additionally, both state- and regional-level regulations advocate for long-term fire risk assessment. Spanish forestry legislation (specifically, Law 43/2003) states that each regional ministry of the environment must declare areas of high fire risk as those with extra fire frequency or additional severity, and must adopt special fire prevention measures, while a main goal of the regional civil protection plan for forest fire emergencies in Castilla y León (INFOCAL) is zoning according to long-term fire risk arising from natural and human causes.

To date, parametric regression techniques such as generalized linear modelling (GLM) have been widely used to explore critical factors involved in lightning-caused fires and to predict fire occurrence (Chuvieco et al., 2010; Nieto, Aguado, García, & Chuvieco, 2012; Pacheco, Aguado, & Nieto, 2009; Vecín-Arias, Castedo-Dorado, Ordóñez, & Rodríguez-Pérez, 2016). Advances in computer-assisted statistical analysis techniques allow other statistical methods to be more easily implemented, such as random forest models (RFM) (Arpaci, Malowerschnig, Sass, & Vacik, 2014; Guo et al., 2016; Oliveira et al., 2012; Satir, Berberoglu & Donmez, 2016; Vecín-Arias et al., 2016), generalized additive models (GAM) (Brillinger, Preisler, & Benoit, 2003; Brillinger, Preisler, & Benoit, 2006; Vilar, Woolford, Martell, & Martín, 2010; Woolford et al., 2016) and spatial autoregressive models (AUREG) (Beron & Vijverberg, 2004; LeSage, 2000; McMillen, 1992; Pace & Barry, 1997).

Since spatial correlation of observations plays an important role in explaining lightning-caused fire occurrence, it is important to define suitable strategies that include these. One way is to explicitly include planar coordinates as covariates in the models. Another way – in GAM-based methods – is to include a non-parametric spatial trend function with planar coordinates as

arguments in the model (Woolford et al., 2011). Alternatively, spatial autoregressive models can tackle the issue of spatial correlation by including not only the values of the dependent variable for each observation but also for the surrounding area. This is normally accomplished by means of spatial weight matrices that contain information on the spatial relationship between observations (Martinetti & Geniaux, 2017; Wilhelm & Godinho de Matos, 2013).

Although GAM-based methods that include a spatial trend function (spatial GAM, abbreviated hereinafter as GAMs) have previously been used for the evaluation of wildfire occurrence (Brillinger et al., 2003; Brillinger et al., 2006; Vilar et al., 2010; Woolford et al., 2011) and for lightning-caused wildfire forecasting (Read, Duff, & Taylor, 2018), this approach has not been used to model and predict lightning-caused fires.

Our aim was to demonstrate the usefulness of GAMs in modelling and predicting long-term lighting-caused fire risk at the regional scale in Castilla y León. Theories and models regarding the main factors affecting lightning-caused fires reveal that the relative importance of these factors can vary according to the studied area (Krawchuck et al., 2006; Little, McKenzie, Peterson, & Westerling, 2009), for which reason the development of models at the regional scale is advisable (Collins, Price, & Penman, 2015; Nieto et al., 2012; Pachecho et al., 2009; Reineking, Weibel, Conedera, & Bugmann, 2010).

We compared GAMs predictions that included spatial effects with GAM, GLM and RFM predictions, which consider longitude and latitude as independent covariates, and with AUREG predictions that take into account spatial correlation through spatial weight matrices. Alternative methods such as support-vector machines, naïve Bayes classifiers and artificial neural network algorithms could also be applied to binary response data. However, our selected models are representative of the state-of-the-art in regression for binary responses and so are appropriate to the purpose of this research. In reporting our findings, we also describe the

practical usefulness of GAMs predictions for fuel and fire management in the Castilla y León region.

2. MATERIAL AND METHODS

2.1. Study area

The studied area was Castilla y León, an autonomously governed region of Spain located in the centre-north of the Iberian Peninsula (Fig. 1). The region mainly consists of a plateau surrounded by several mountain chains (highest peak 2,648 m and mean altitude 830 m). Mean annual rainfall, conditioned by the orography, varies between 1,000 mm for the northern mountain ranges and 400 mm for the plateau (Nafría et al., 2013). Average annual temperature is around 11°C. Forests (including woodland and shrubland) cover over half the region (51%), cropland accounts for 31.4% (mainly the central plateau) and the remaining 17.6% is pastureland for extensive livestock farming. The predominant woodland species are *Quercus ilex* L. (15.1% of the region), *Quercus pyrenaica* Willd. (15.0%), *Pinus pinaster* Aiton (8.6%) and *Pinus sylvestris* L. (7.0%) (Consejería de Medio Ambiente, 2005).

2.2. Data sources

Data on lightning, rainfall, land cover, topography and forest fire ignitions were sourced from several state bodies in Spain.

Lightning data was provided by the Spanish Meteorological Agency (AEMET), whose lightning detection network (LDN) includes 15 lightning sensors in Spain and 4 in Portugal (Fig. 1). These IMPACT (Improved Performance from Combined Technology) sensors, equipped with temporal GPS technology, form part of other worldwide LDNs (Pérez-Puebla, 2004). The LDN detects and locates ground-strike locations of cloud-to-ground lightning flashes, but does not provide data on intra-cloud discharges. For the entire Iberian Peninsula, it

was reported that the detection efficiency of the LDN (the probability of detection) (Nagh, Murphy, Schulz, & Cumminss, 2005) is currently greater than 90% and that the median lightning location error is around 0.5 km (Pérez-Puebla, 2004). Lightning flashes are characterized in terms of the total number of strikes, strike duration, strike polarity, current peak for both positive and negative strikes, location (longitude and latitude) and estimated error ellipse.

Daily rainfall data was obtained from 395 AEMET weather stations and 57 Agricultural Technological Institute of Castilla y León (ITACyL) weather stations (a total of 452 stations). Detailed information on land cover was obtained from the Spanish digital forestry map of the studied area (Ministerio de Medio Ambiente, 2003), in which the minimum plot size is 2.25 ha for forested areas and 6.25 ha for other land uses (Robla-González, Vallejo-Bombín, De La Cita-Benito, & Lerner-Cuzzi, 2009). Topographic data (altitude, slope and aspect) were obtained from a digital terrain model with a resolution size of 200 × 200 m, provided by the Spanish National Geographic Institute (IGN). Finally, forest fire data was obtained from the Spanish Ministry of Agriculture, Fishing, Food and the Environment (MAPAMA), which provides information on ignition point coordinates, the date and time of detection, the cause of the ignition and the surface area burned by each fire. The database contains data on all forest fires that occurred in Spain, regardless of their final size.

2.3. Data pre-processing

The study covered the five months of May to September for the 11 years 2000-2010, during which a total of 662 lightning-caused fires with available planar coordinates were reported for Castilla y León. Most of the thunderstorms in this region occur between May and September, when temperatures foster the development of convection processes (Rivas-Soriano, de Pablo, & Tomas, 2005). In consequence, the vast majority of lightning-caused wildfires in Castilla y León are recorded during that extended summertime period (Vecín-Arias et al., 2016).

Fig. 2 shows the number and distribution of lightning-caused fires across size classes. Most of the fires burned less than 1 ha (i.e., they were extinguished almost as soon as they began) and fire-burned areas size measured on average between 0.01 ha and 1303.83 ha.

Using the ArcGIS software (v10.2; ESRI Inc., Redlands, CA, USA) all available lightning, rainfall, land cover, topography and forest fire data were georeferenced in a map grid with a spatial resolution of 4 × 4 km (yielding 6,253 grid cells). This grid, the finest resolution advisable with the currently available data, was mainly limited by interpolated rainfall and lightning location error. The distribution of the 662 reported lightning-caused fires within this grid is shown in Fig. 3. In the studied period, only one wildfire occurred in 452 cells, whereas 2 wildfires occurred in 75 cells, 3 wildfires occurred in 16 cells and 4 wildfires occurred in just 3 cells.

Fig. 3 also shows the spatial distribution of flash density (flashes km^{-2} $year^{-1}$) in Castilla y León for the period 2000-2010. A total of 533,173 flashes were considered. The number of thunderstorm days (days in which at least one flash was recorded) and dry thunderstorm days (thunderstorm days with daily rainfall of 2.5 mm or less) (Álvarez-Lamata, 2005; Rorig & Ferguson, 1999) were counted for all the grid cells in the map. To calculate daily rainfall for each grid cell, a continuous rainfall map was computed by spatially interpolating daily rainfall data from the 452 weather stations using a simple co-kriging methodology, which calculates rainfall for each grid cell using a multivariate spatial model, and elevation as a related secondary attribute (Carrera-Hernández & Gaskin, 2007; Jarvis & Stuart, 2001).

Aspect and slope were derived from the digital terrain model using standard algorithms (Burrough & McDonnell, 1998). To be able to deal with more homogeneous types of land cover, Spanish forestry map information was reclassified in the following 7 groups: coniferous (patches dominated by pure coniferous stands), broadleaf (patches dominated by pure broadleaf

stands), mixed (patches with a mixture of coniferous and broadleaf trees), shrubland, cropland, pastureland and non-combustible areas (urban areas, arid areas, rock formations, wetlands, etc). Table I shows the list of variables computed for each 4 × 4-km grid cell.

2.4. Model fitting

We compared five different models: GLM, RFM, GAM, GAMs and AUREG. As mentioned above, GAMs refer to a GAM that includes a spatial trend function, and AUREG to a spatial probit autoregressive model. As our aim was to demonstrate the usefulness of the GAM-based approach to modelling lightning-caused fire occurrence, we will only describe this technique in detail.

A GAM is a non-parametric generalized linear model with a linear predictor involving a sum of smooth functions of covariates (Wood, 2006). These smooth functions are used as a link function to set up an additive relationship between the mean response and the covariates (Guisan, Edwards, & Hastie, 2002; Wood & Augustin, 2002).

In our study, the binary response variable Y was defined as the occurrence of at least one lightning-caused fire within each grid cell in the study period; i.e., lightning fire occurrence was modelled in binary form (the absence or presence of fires, coded 0 and 1, respectively). A binary dependent variable (presence/absence) is justified given that lightning-caused fires are rare events in the region: more than one fire occurred in only a third (31.7%) of the cells where lightning-fires occurred over the 11-year study period.

The potential predictors or covariates $X = (X_1, \ldots, X_p)$ are those described in Table I. In spatial data, the variables (X, Y) have a spatial location given by $s = (s_1, s_2)$, where s_1 denotes longitude and s_2 denotes latitude. In our particular case, s_1 and s_2 are given in projected UTM-30N coordinates (reference system ETRS89).

Since the dependent variable followed a binomial distribution, we used a logistic GAM model with a binary response given by:

$$P(Y=1\,|\,\boldsymbol{s},\boldsymbol{X})=\frac{exp\left(\alpha+g\left(s_1,s_2\right)+f_1\left(X_1\right)+_{...}+f_p\left(X_p\right)\right)}{1+\exp\left(\alpha+g\left(s_1,s_2\right)+f_1\left(X_1\right)+_{...}+f_p\left(X_p\right)\right)} \quad (1)$$

where α is a constant, $g(\cdot)$ is a smooth function of bivariate spatial effect, and $f_j(\cdot)$ $(j = 1, \ldots, p)$ are smoothing splines that account for non-linear relationships between the probability of a lightning-caused fire occurrence and the explanatory variables X_j. Note that identification is guaranteed by introducing the constant α into the model, while zero mean is required for the spatial effect g and for the partial functions $f_j(\cdot)$(Hastie & Tibshirani, 1990).

The model represented by Eq. (1) is referred to as GAMs, which differs from a GAM in that it includes planar coordinates together as model inputs through the function g (s_1, s_2). Both GAM-based models were fitted with the `gam` function in the `mgcv` package from the statistical environment `R` (R Development Core Team, 2016). Using this function, and given a sample $\{s_i, X_i, Y_i\}_{i=1}^{n}$, where n is the sample size, we were able to estimate the components $\hat{\alpha}$, $\hat{g}$, $\hat{f}_j$ of the model in Eq. (1).

Functions $g(\cdot)$ and $f_j(\cdot)$ were estimated using cubic splines smoothers (Hastie & Tibshirani, 1986) of the form:

$$S\left(X\right)=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\sum_{j=1}^{k}\lambda_j\left(X-\xi_j\right)_+^3$$

where (2)

$$\left(X-\xi_j\right)_+=\begin{cases}X-\xi_j, X>\xi_j\\ 0, \text{otherwise}\end{cases}$$

with $\xi_j, j=1,...,k$ as the knots, i.e., points where the cubic polynomials join.

The bivariate term g (s_1, s_2) in Eq. (1) is approximated using 2D splines that can be described in a similar way as in Eq. (2), except that a 16-coefficient cubic polynomial is used with terms of the type $X^{\alpha}Y^{\beta}$, with $\alpha=0,1,..,3$ and $\beta=0,1,..,3$.

GLM models were fitted with the `glm` function from the statistical environment `R`. For the RFM and AUREG models we used the `ranger` and `ProbitSpatial` `R` packages, respectively. GLM and GAM models follow a logistic structure similar to that reflected in Eq. (1) for the GAMs model. RFM is an assembly of decision trees that follows a different strategy that consists of recursively dividing the *p*-dimensional space into regions according to rules aimed at selecting the most informative variables. AUREG follows a spatial autoregressive probit model, which is comparable to the logistic model, although the link function is the inverse of the standard normal distribution instead of the logistic function. The mathematical expression for a spatial autoregressive model is:

$$Y = \rho \mathbf{W} Y + \beta_1 X_1 + ... + \beta_p X_P + \varepsilon = \rho \mathbf{W} Y + \mathbf{X}\boldsymbol{\beta} + \varepsilon \quad (3)$$

where $\mathbf{W}\ (n \times n)$, n – the number of observations – is a weight matrix that captures the spatial dependence between neighbouring observations. $\mathbf{W}$ values are zero in the diagonal, and rows are normalized weights of the form $w_{ij} = 1/m_i$, where m_i is the number of observations contiguous to *i*. Elements in the i^{th} row of $\mathbf{W}$ not defined as contiguous to observation *i* are 0. ρ, with $abs(\rho) < 1$, is a scalar parameter that measures the strength of the spatial dependence ($\rho = 0$ means no spatial dependence). $\varepsilon : N\ (0, \sigma^2)$ is the error term.

In the Bayesian approach to autoregressive models, *Y* in Eq. (3) is replaced by an unobserved latent variable Z related to the observable binary dependent variable as follows: Y ($Y = 1$, if $Z \geq 0$; $Y = 0$, if $Z < 0$). The idea is to decompose the posterior distribution of the parameters, given the data *Y*, $p\left(Z, \beta, \rho \middle| Y\right)$, into a set of conditional distributions for each parameter in the model. A Markov chain Monte Carlo (MCMC) scheme is usually used to estimate the posterior drawing samples from prior distributions of Z, β and *Y* (Wilhelm & Godinho de Matos, 2013). MCMC assumes that the form of a probability density can be

approximated from a large sample using kernel density estimators or histograms (LeSage & Pace, 2009).

2.5.Prediction evaluation

We used receiver operating characteristic (ROC) curves (Swets & Pickett, 1982) to estimate the probability of false positives and false negatives in the predictions made by the five tested models. The ROC curve is a plot, for all possible cut-points, of sensitivity ($Sens$, the instances correctly classified as positive as a percentage of the total number of true positives) versus 1-specificity ($Spec$, the instances correctly classified as negative as a percentage of the total number of true negatives). From the ROC curves we calculated the *F* score statistic in order to obtain a threshold (cut-point) that yielded a measure of the effectiveness of the predictions for unbalanced data (such as those that concern us here) (Van Rijsbergen, 1979). *F* score was selected because it is less influenced by unbalanced data, which is the case here. The general formula of the *F* score statistic for positive real δ is:

$$F_\delta = \left(1 + \delta^2\right) \cdot \frac{Prec \cdot Sen}{\left(\delta^2 \cdot Prec\right) + Sen} \quad (4)$$

where *Prec* (precision) is defined as the ratio between instances correctly classified as positives and all the positive classified instances (whether or not correctly classified).

We used δ=1 and hence, the *F*1 score was calculated as $F1 = 2\frac{Prec \cdot Sen}{Prec + Sen}$, i.e., the harmonic mean between precision and sensitivity. The larger the *F*1 score, the better the prediction, taking into account both precision and sensitivity. Although we gave the same weight to precision and the sensitivity in calculating *F*1 (we set $\delta = 1$), it would also be possible to assign different weights to each just by changing the δ value.

Finally, using the bootstrapping technique (random sampling with replacement), we estimated the sample distributions for the *F*1 score and plotted these distributions using boxplots. In each repetition, 70% of the data (4,377 grids) was used for model fitting and the remaining 30%

(1,876 grids) was used to calculate the $F1$ score. Accordingly, as many models were fitted as bootstrap repetitions were implemented.

2.6. Variable selection

The goal with variable selection was to determine the best subset of q $(q \leq p)$ covariates to include in the model to ensure the best predictive capacity. A larger number of covariates does not necessarily lead to better models for two main reasons. Firstly, the resulting models are difficult to interpret, as well as being prone to collinearity and overfitting. Secondly, because of the bias-variance trade-off, the inclusion of irrelevant covariates would increase the variance of the estimated coefficients, resulting in a loss of predictive ability (higher variability in a prediction for any given data point).

We opted for a two-stage variable selection algorithm to select the best prediction model. The best combination of q variables was first selected using a step-by-step procedure and the number of covariates to be included in the model was then determined.

In the first step, given a number $q(q \leq p)$ of covariates, the objective was to find the combination of q variables that provides the maximum $F1$ score. Let $F1_{j1,\dots,jq}(j_1 < j_2 < \cdots < j_q)$ be the $F1$ score (computed as explained above), obtained using only q covariates and leaving out the remaining variables. We use $j_1, \dots, j_q$ as subscripts, instead of $1, \dots, q$, to indicate that the order does not have to be consecutive. Based on this metric, the best q predictors $X_{j1}, \dots, X_{jq}$ could then be selected. The vector of indices $(l_1, \dots, l_q)$, which indicates which covariates are included in each model of q covariates, was obtained by maximizing:

$$(l_1, \dots, l_q) = \underset{j_1,\dots,j_q}{\operatorname{argmax}} F1_{j_1,\dots,j_q} \tag{5}$$

For simplicity sake, we denote the value of the $F1$ score obtained for the combination of q variables as $F1(q) = F1_{l_1,\dots,l_q}$, while, $l_1, \dots, l_q$ are the indices of those variables that maximize

the $F1$ score. Once the best combination of q predictors was obtained, the optimal number of covariates to be included in the model was selected by setting the value for q that maximized $F1(q)$. In order to shorten the process, we did not test all the models for each combination of q=1,...,p variables, but limited the study to a maximum of q = 6 (as will be seen, fewer variables result in the most appropriate models).

Finally, the model selected was not that with the maximum $F1$ score, but the model that, having a $F1$ score statistically comparable to the maximum, has the minimum number of predictors. This was done by determining a confidence interval of the type $[a,\infty]$ for the difference $F1(q) - F1(q-1)$ by means of bootstrapping. The simplest model (that with the minimum value of q) that was statistically comparable to other more complex models (with a greater value of q) was that corresponding to the maximum value of q for $a > 0$.

In order to evaluate how clustered or dispersed the errors were for the tested methods, the spatial correlation of errors (misclassified grid cells) was evaluated by computing the average nearest-neighbour distance index (ANNDI), which reflects the ratio of the observed mean distance to the expected mean distance (average distance between neighbours in a hypothetical random distribution). It can be proved that the nearest neighbour distance between observations follows a Rayleigh distribution, which is closely related to the normal distribution (Smith, 2016). It is also easy to prove that the expected mean distance is given by $D_e = \frac{0.5}{\sqrt{N/A}}$, where N is the number of observations and A is the area of the minimum rectangle enclosing all the cells (i.e., the denominator represents the point density) (Smith, 2016). If ANNDI is less than or greater than 1, then the pattern trends toward clustering or dispersion, respectively. The interval range is 0 to 2.14.

3. RESULTS AND DISCUSSION

Table II shows the maximum $F1$ scores for the different models and predictors (q =1 to 6) and the corresponding sensitivity, specificity and *CCR* (the correct classification rate). For the same number of predictors, GAMs outperformed GLM, GAM and RFM (see also Fig. 4, which shows the larger $F1$ scores obtained for the GAMs, with $q = 1$ to 6 covariates for the test dataset). At this point, it must be noted that GAMs starts from a model with s_1 and s_2 as covariates, so strictly speaking, for this specific model, q equals q+2 in Table II. The *CCR* values obtained compare favourably with those for other models in the lightning-caused wildfire occurrence literature in Spain (Castedo-Dorado, Rodríguez-Pérez, Marcos-Menéndez, & Álvarez-Taboada, 2011; Chuvieco et al., 2010; Pacheco et al., 2009; Vecín-Arias et al., 2016).

According to the $F1$ scores, the top-performing model was GAMs with five covariates, namely, the percentage of coniferous woodlands (%coniferous), the percentage of agricultural crops (%crops), the percentage of north-facing aspects (%North), the percentage of mixed forests (%mixed), and the percentage of non-combustible areas, i.e., urban, mining, water, waste disposal sites, etc. (%other). For the 30 combinations of methods and covariates tested, %coniferous and %crops were the most common covariates, being present in 29 and 18 combinations, respectively.

Fig. 5 shows the partial dependence plots for the explanatory variables included in the GAMs for q = 1 to 5. The size of the confidence intervals is associated with the number of observations (note how the width of the confidence intervals increases as the number of observations of the covariates decreases). Although, for the same covariate, slight differences can be observed in the trends of the partial dependence plots depending on q, a consistent pattern is evident regarding each influencing factor across all the models.

According to Fig. 5, the percentage of coniferous woodlands is positively associated with the probability of lightning-caused fires; furthermore, this relationship is non-linear, so a non-

parametric model is justified. The probability of lightning-caused forest fires also increases faster and more precisely for low values of this covariate. Although lightning ignition is not a fully understood process, most studies indicate that the object of strikes is generally a tree (Larjavaara, Pennanen, & Tuomi, 2005; Mäkelä, Karvinen, Porjo, Mäkelä, & Tuomi, 2009), with fire ignition typically occurring on the ground (in litter, duff, mosses, etc) in sheltered sites near the tree bole that acts as the conductor of the lightning discharge (Larjavaara et al., 2005; Latham & Schlieter, 1989). Duff and litter favour fire ignition, while understory shrub species or fine woody debris favour fire spread (Mäkelä et al., 2009; Latham & Schlieter, 1989). This fact could explain why lightning-induced fires affect a greater proportion of woodlands than human-induced fires (Vázquez & Moreno, 1998).

For temperate European countries lightning fires are more likely to occur in coniferous and mixed coniferous forests, as reported for Switzerland (Conedera, Cesti, Pezzatti, Zumbrunnen, & Spideni, 2006; Pezzatti, Bajocco, Torriani, & Conedera, 2009; Reineking et al., 2010) and Austria (Müller et al.,2013; Vacik & Müller, 2017). In contrast, fires caused by lightning tend to be underrepresented in broadleaf stands such as those dominated by chestnut, oak or beech trees. The greater probability of lightning-caused fires associated with conifer stands has also been reported in Spain (Castedo-Dorado et al., 2011; Nieto et al., 2012; Pineda, Montanyà, & van der Velde, 2014). According to some authors the thicker litter and duff layer and the higher flammability associated with coniferous species compared to broadleaf species are postulated to be key causes (Conedera et al., 2006; Flannigan & Wotton, 1991; Latham & Williams, 2001). Additionally, an abundance of highly flammable species in the understory of conifer stands of *Erica* sp., *Genistella tridentata*, *Calluna vulgaris*, etc. may aid fire propagation after ignition, contrary to what usually occurs in broadleaf forests (Bond & Van Wilgen, 1996).

The moderately positive effect of the percentage of mixed woodlands on lightning fire occurrence is probably related to the presence of coniferous species in these stands (Reineking

et al., 2010). For mixed temperate forests, it has been reported that lightning discharges tend to strike the tallest conifer trees, which act as ground terminals and thus increase the likelihood of fire occurrence (Yanoviak et al., 2015).

The negative effect of the percentage of agricultural crops and non-combustible areas (urban areas, arid areas, rock formations, wetlands, etc) would point to a low probability of lightning fire ignition and spread in areas with a low percentage of forest and natural cover. This result could be expected *a priori*, given the low flammability and combustibility of this type of land cover. In a countrywide study for Spain, the percentage of wildland area at the municipality level has been reported as the most important factor to discriminate non-fire-prone from fire-prone areas (Martínez-Fernández, Chuvieco, & Koutsias, 2012).

The negative effect of the percentage of north-facing aspects (i.e., less likelihood of fire occurrence when the percentage of landscape facing north is high) also seems intuitively comprehensible, as fires occurring on terrain facing suntraps are more likely to spread due to higher solar incidence and comparatively drier fuel (Vankat, 1985; Vasconcelos, Silva, Tomé, Alvim, & Pereira, 2001). This feature has elsewhere been reported to increase ignition likelihood (Vankat, 1985).

Fig. 6 shows boxplots of $F1$ scores for each method and number of covariates (q =1 to 6). To obtain a quasi-steady maximum $F1$ score, visually it would seem that only two covariates were required for all the methods.

These results were statistically tested by bootstrapping a confidence interval of the type $[a, \infty]$ for the difference $a = F1(q) - F1(q-1)$. The best models for q and q-1 were obtained for the same bootstrap samples during training. $F1$ scores were then calculated using the same bootstrap samples from the test dataset. Fig. 7 shows $F1$ score differences for the best model as a solid line, while the lower limit of the confidence interval $[a, \infty]$ for the differences is represented by a dashed line. The minimum number of statistically significant covariates in

each model corresponded to the first value of q before $a < 0$ (i.e., the value of q after the dashed line crossed the red line). Note that GLM, GAM, GAMs and AUREG only needed one covariate to obtain an $F1$ score comparable to that of the other models with $q > 1$. Conversely, RFM required two covariates to obtain the best solution.

Boxplots of $F1$ scores for each method according to the number of covariates are depicted in Fig. 8. Once again, it can be observed that the best results correspond to the GAMs, irrespective of the value of q.

To analyse the statistical significance of these results, we repeated the procedure used to determine the minimum number of covariates. Accordingly, we determined confidence intervals $[a, \infty]$ for the increment in $F1 = F1(\text{model1}) - F1(\text{model2})$, where model1 has q covariates and model2 q-1 covariates. From Fig. 9 it is clear that there was a significant increment in $F1$ scores when the GAMs was compared to the other methods, regardless of the number of covariates. Note that the fact that GAMs includes two covariates (planar coordinates) more than all the other models for the same value of q does not affect the model choice or qualitatively change the interpretation of the results, given that GAMs outperforms the remaining models even when it contains two or more additional covariates.

The GAMs with just one covariate, namely the percentage of coniferous woodlands, was finally selected as the most appropriate model for prediction purposes – as the simplest model with an $F1$ score comparable to other models with $q > 1$. This, the most informative predictor, confirms the importance of fuel in the occurrence of lightning-caused fires, as has been documented elsewhere (Krawchuk et al., 2006; Mundo, Wiegandc, Kanagarajc, & Kitzbergerd, 2013; Vasconcelos et al., 2001).

The fact that, for the same number of covariates, the GAMs was always superior to the GAM confirms that lightning-caused fires depend to some degree on geographic location (Clarke, Gibson, Cirulis, Bradstock, & Penman, 2019; González-Olabarría, Mola-Luyego, & Coll,

2015). The contour plot of the estimated bivariate spatial effect (Fig. 10) suggests that there is a lower risk of fire lightning-caused fire occurrence towards the centre of the region, and a higher fire risk at the boundaries, primarily in the south and the east. Thus, the spatial effect roughly matches the reported distribution of fires (Fig. 3) and provides additional information not recorded by the covariate included in the selected model. The fact that the geographic location plays an important role in lightning-induced wildfires suggest that region-specific data collection and modelling, as described in this paper, need to be prioritized in future works.

Fig. 11a depicts the spatial distribution of predicted lightning-caused fires using GAMs probability estimates and considering a cut-point of 0.15 (the point on the ROC curve where the $F1$ score is maximum), showing correctly and incorrectly classified fires. For this cut-point, the GAMs estimated a higher percentage of false positives than of false negatives (Fig.11c). This behaviour, which reflects an overestimation of fire occurrence, is common to most models developed to predict lightning-caused fires (Castedo-Dorado et al., 2011; Nieto et al., 2012; Pacheco et al., 2009). We consider that this behaviour, reflecting a tendency to err on the side of caution in predicting areas where lightning-caused fires are likely to occur, is permissible given the purposes of our study (i.e., to ultimately reduce population and economic risk posed by wildfires).

Estimated linear coefficients, their significance values and estimated significance of the smoothed terms in the selected GAMs are shown in Table III. All terms were highly statistically significant.

Regarding the spatial distribution of model errors, the ANNDI analysis indicated that misclassified grid cells (Fig. 11c) were clustered (ANNDI = 0.785; $p < 0.0001$). This result would suggest that the GAMs does not fully incorporate a local spatial effect associated with the $g(\cdot)$ term in Eq. (1). Note that not being able to remove the spatial correlation in the residuals of the model may affect estimates of model coefficients and their respective p-values (Table

III). Nevertheless, considering that the *p*-values are very close to zero, it is not expected that the spatial correlation will affect the statistical significance of the coefficients.

4.1. Management implications

Understanding where wildfires are most likely to occur is critical to determining where wildfires pose the greatest risk to people and property. From an operational point of view, a better knowledge of the spatial patterns of lightning fire occurrence and their relationships with underlying risk factors is necessary to ensuring that prevention efforts are more efficient.

Long-term lightning-caused fire risk evaluation can inform the zoning of a region according to proneness to lightning fires, as required by INFOCAL for the autonomously governed Castilla y León region. Operationally, risk assessment can help regional ministry of the environment in several ways, as follows: (i) in identifying areas where wildfire detection based on lookout towers is not effective, (ii) in making informed decisions regarding preparedness planning and fuel management, and (iii) in designing strategically rational firefighting responses.

The regional government of Castilla y León is currently responsible for a network of 198 lookout towers in the region (https://medioambiente.jcyl.es), strategically located so as to detect human-induced wildfires, more than wildfires of natural origin. Thus, many lightning-fire-prone areas are not directly visible from the lookout towers, especially in the northwest and south of the region (Fig. 12). Better identification of fire prone areas should lead to a reconsideration of lookout tower locations or the instigation of ground patrols during the wildfire season so as to ensure earlier fire detection (Kucuk, Topaloglu, Altunel, & Cetin, 2017).

Fuel management is considered crucial to reducing wildfire spread and severity in Mediterranean areas, especially of larger wildfires (Moreira et al., 2011). Fuel management is usually implemented through fuel modification and fuel type conversion strategies (Rigolot, Fernandes, & Rego, 2009). Our results suggest that fuel modification efforts should

strategically focus on coniferous woodlands, given our finding that this kind of forested land is primarily associated with lightning-caused fires. As mentioned above, probable reasons for the greater likelihood of fire occurrence in conifer forests are forest-floor fuel (the litter and duff layer) and the flammability of conifers and the associated understory species.

Fuel management can be addressed by surface fuel-modification options, including prescribed burning, shrub clearing and understory mastication, aimed at reducing lightning-caused fire ignition and spread risks. Prescribed burning of pine stands tolerant of low-intensity fire is an effective way to reduce the fuel load and so diminish the risk of ignition and spread (Kucuk et al., 2017). While shrub clearing and mastication of understory growth do not reduce the probability of fire ignition (the litter load is not reduced and mastication actually increases litter depth), they do induce changes in fire spread patterns (Fernandes et al. 2013). Crown fuel modification options that address the spread of lightning-caused fires include thinning and pruning operations.

Fuel type conversion in wildland areas would involve replacing coniferous stands by a mixture of coniferous and broadleaved species, e.g., selective thinning in favour of broadleaved species or forest management that promotes natural succession. Nevertheless, according to our results, fuel conversion would not be a suitable solution, as a higher likelihood of fire occurrence is also associated with a large percentage of mixed forests in the landscape.

Our results also confirm that forested areas are more fire-prone than agricultural areas, suggesting that permanent croplands and pastures located within a more heterogeneous landscape would result in fewer fire ignitions, with the farmlands acting as fire breaks that enhance resistance to fire spread (Vega-García & Chuvieco, 2009).

Since the regional ministry of the environment of Castilla y León is directly responsible for preventing fires in publicly managed forests – representing 40% (around 20,000 km^2) of forested surface in the region (Junta de Castilla y León, 2009) – from an operational viewpoint

forest managers need to include the reduction of lightning fire risk as an additional objective in forest management planning. Fire risk mapping, as proposed in our study, would identify the spatial foci for long-term prevention actions, including curtailment of understory biomass load and crown fuel load and the promotion of diversified landscapes.

Suppression (or initial attack) models attempt to model the ability of firefighting resources to contain a fire before it goes out of control and before damage is caused to people or property (Plucinski, 2013). Risk assessment models can inform the design and development of fire management plans and responses to wildfires. By including information on fire-prone locations in fire behaviour simulators, fire spread patterns can be predicted in order to plan how firefighting resources are deployed with a view to reducing risk and containing spread (Bahro, Barber, Sherlock, & Yasuda, 2007; Costa, Castellnou, Larrañaga, Miralles, & Kraus, 2011). This approach would be especially important for Castilla y León, given its high wildfire occurrence rate and accounting for 20% of the total surface burned in Spain in the period 2001-2010 (MAGRAMA, 2012). If fire load is high (e.g., as happened during the 2017 fire season), temporal constraints in suppression activities can be improved through effective analysis of fire behaviour and optimization of resource deployment, to the point of even leaving lower priority fires watched but unattended.

4. CONCLUSIONS

Using five alternative models, namely GLM, RFM, GAM, AUREG and GAMs, we modelled the relationship between lightning-caused fires and a set of potential covariates for a case study referring to Castilla y León region in central Spain. Taking into account both predictive ability and simplicity, we found that the most suitable model was a GAMs that considered a single covariate, namely, percentage of coniferous woodlands in the landscape, along with the planar coordinates (whose effects were modelled using a bivariate function). This finding reflected the

following results: (i) that lightning fire occurrence is more closely related with vegetation type than with topography, climate or lightning characteristics, and (ii) that it was important to take account of the location of reported lightning fires through a spatial bivariate function (i.e., not as separate covariate). Additionally, in terms of evaluating lightning-caused fire occurrence, our study points to the advantages of using GAMs over other statistical techniques whose use has predominated in the historical fire risk evaluation literature to date. The model selected is intended to better inform and potentially improve wildfire management by identifying areas where wildfire detection through lookout towers is not effective and aiding informed decision-making regarding preparedness planning and fuel management and the design of strategically rational firefighting responses.

Although the results obtained in our case study cannot be considered as directly transferable to other cases, what can be generalized to other regional analyses is the described GAMs and variable selection methodology.

Acknowledgments

Funding for this research was provided by the Universidad de León under the project titled "Análisis de la distribución espacial y temporal y caracterización de fenómenos tormentosos en el medio agrícola y forestal de la Meseta Central y del Norte de España". AEMET provided the lightning strike and meteorological data used.

REFERENCES

Álvarez-Lamata, E. (2005). Los incendios forestales y las condiciones meteorológicas en Aragón. In SECF & Gobierno de Aragón (Eds.), *Proceedings of 4º Congreso Forestal Español*. Zaragoza (Spain).

Arpaci, A., Malowerschnig, B., Sass, O., & Vacik, H. (2014). Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. *Applied Geography, 53,* 258-270.

Bahro, B., Barber, K. H., Sherlock, J. W., & Yasuda, D. A. (2007). Stewardship and Fireshed Assessment: A Process for Designing a Landscape Fuel Treatment Strategy. In P.F. Power (Ed.). *Restoring fire-adapted ecosystems: proceedings of the 2005 national silviculture workshop, USDA Forest Service Gen. Tech. Rep. PSW-GTR-203* (pp. 41-54). Albany, CA.

Beron, K. J., & Vijverberg, W. P. M. (2004). Probit in a spatial context: a Monte Carlo analysis. In L. Anselin, R. J. G. M. Florax & S. J. Rey (Eds.), *Advances in Spatial Econometrics. Methodology, Tools and Applications* (pp. 169-195). Berlin, Germany: Springer.

Bond, W. J., & Van Wilgen, B. W. (1996). Why and how do ecosystems burn?. In W.J. Bond & B.W. Van Wilgen (Eds.), *Fire and Plants* (pp.17–33). London: Chapman and Hall.

Brillinger, D. R., Preisler, H. K., & Benoit, J. W. (2003). Risk assessment: a forest fire example. *Institute of Mathematical Statistics Lecture Notes, 40,* 177-196.

Brillinger, D. R., Preisler, H. K., & Benoit, J. W. (2006). Probabilistic risk assessment for wildfires. *Environmetrics, 17*, 622-633.

Burrough, P. A., & McDonnell, R. A. (1998). *Principles of Geographical Information Systems*. London, UK: Oxford University Press.

Carrera-Hernández, J. J., & Gaskin, S. J. (2007). Spatiotemporal analysis of daily precipitation and temperature in the Basin of Mexico. *Journal of Hydrology, 336,* 231-249.

Castedo-Dorado, F., Rodríguez-Pérez, J. R., Marcos-Menéndez, J. L., & Álvarez-Taboada, M. F. (2011). Modelling the probability of lightning-induced forest fire occurrence in the province of León (NW Spain). *Forest Systems, 20*, 95-107.

Chow, J. Y. J., & Regan, A. C. (2011). Resource location and relocation models with rolling horizon forecasting for wildland fire planning. *Infor: Information Systems and Operational Research*, *49*, 31–43.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M. P., Vilar, S., Martínez, J., Martín, S., Ibarra, P., de la Riva, J., Baeza, J., Rodríguez, F., Molina, J. R., Herrera, M.A., & Zamora, R. (2010). Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. *Ecological Modelling*, *221,* 46–58.

Chuvieco, E., Salas, F. J., Carvacho, L., & Rodríguez-Silva, F. (1999). Integrated fire risk mapping. In E. Chuvieco (Ed.), *Remote sensing of large wildfires in the European Mediterranean Basin* (pp. 61-84). Berlin, Germany: Springer-Verlag.

Clarke, H., Gibson, R., Cirulis, B., Bradstock, R. A., & Penman, T. D. (2019). Developing and testing models of the drivers of anthropogenic and lightning-caused wildfire ignitions in south-eastern Australia. *Journal of Environmental Management, 235*, 34–41.

Collins, K. M., Price, O. F., & Penman, T. D. (2015). Spatial patterns of wildfire ignitions in south-eastern Australia. *International Journal of Wildland Fire, 24,* 1098–1108.

Conedera, M., Cesti, G., Pezzatti, G. B., Zumbrunnen, T., & Spinedi, F. (2006). Lightning-induced fires in the Alpine region: An increasing problem. In D. X. Viegas (Ed.) *Fifth International Conference on Forest Fire Research* (9 pp.). Coimbra, Portugal: ADAI/CEIF University of Coimbra.

Consejería de Medio Ambiente. (2005). *Castilla y León crece con el bosque*. Valladolid (Spain): Junta de Castilla y León.

Costa, P., Castellnou, M., Larrañaga, A., Miralles, M. & Kraus, D. (2011). *Prevention of large wildfires using the Fire Type concept*. Barcelona, Spain: EU Fire Paradox Publication.

Fernandes, P. M., Davies, G. M., Ascoli, D., Fernández, C., Moreira, F., Rigolot, F., Stoof, C. R., Vega, J. A., & Molina, D. (2013). Prescribed burning in southern Europe: developing fire management in a dynamic landscape. *Frontiers in Ecology and the Environment, 11 (Online Issue 1),* e4–e14.

Flannigan, M. D., & Wotton, B. M. (1991). Lightning-ignited forest fires in northwestern Ontario. *Canadian Journal of Forest Research*, *21*, 277–287.

González-Olabarría, J.R., Mola-Luyego, B., & Coll, L. (2015) Different factors for different causes: Analysis of the spatial aggregations of fire ignitions in Catalonia (Spain). *Risk Analysis, 35,* 1197–1209.

Guisan, A. Edwards, T. C. Jr., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling, 157*, 89-100.

Guo, F., Zhang, L., Jin, S., Tigabu, M., Su, Z., & Wang, W. (2016). Modeling anthropogenic fire occurrence in the boreal forest of China using logistic regression and random forests. *Forests, 7*(11)*,* 250.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science, 3,* 297- 310.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models. Monographs on Statistics an Apply Probability*. Boca Raton, FL: Chapman & Hall/CRC.

Jarvis, C. H., & Stuart, N. A. (2001). Comparison among strategies for interpolating maximum and minimum daily air temperatures: part II: the interaction between number of guiding variables and the type of interpolation method. *Journal of Applied Meteorology, 40,* 1075– 1084.

Junta de Castilla y León (2009). *Forests: sign of life in Castile and León*. Retrieved from https://www.foresteurope.org/sites/default/files/LIBRO_SenaldeVIDA_AA_ingles_CERRADO.pdf.

Krawchuk, M. A., Cumming, S. G., Flannigan, M. D., & Wein, R. W. (2006). Biotic and abiotic regulation of lightning fire initiation in the mixedwood boreal forest. *Ecology*, *87*, 458–468.

Kucuk, O., Topaloglu, O., Altunel, A. O., & Cetin, M. (2017). Visibility analysis of fire lookout towers in the Boyabat State Forest Enterprise in Turkey. *Environmental Monitoring and Assessment, 189*, 329.

Larjavaara, M., Pennanen, J., & Tuomi, T. J. (2005). Lightning that ignites forest fires in Finland. *Agricultural and Forest Meteorology, 132*, 171–180.

Latham, D., & Schlieter, J. A. (1989). Ignition Probabilities of Wildland Fuels Based on Simulated Lightning Discharges. In *USDA Forest Service Research Paper INT-411*. Ogden, UT.

Latham, D., & Williams, E. (2001). Lightning and forest fires. In E.A. Johnson & K. Miyanishe (Eds.), Forest Fires: *Behavior and Ecological Effects* (pp. 375–418). San diego, CA: Academic Press.

LeSage, J. P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis, 32*, 19-35.

LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

Littell, J. S., McKenzie, D., Peterson, D. L., & Westerling, A. L. (2009). Climate and wildfire area burned in western US ecoprovinces, 1916–2003. *Ecological Applications, 19,* 1003–1021.

MAGRAMA. (2012). *Los incendios forestales en España. Decenio 2000–2010*. Área de Defensa contra Incendios Forestales (ADCIF). Ministerio de Agricultura, Alimentación y Medio Ambiente (Spain). Report No.: 280-12-210-8.

Mäkelä, J., Karvinen, E., Porjo, N., Mäkelä, A., & Tuomi, T. (2009). Attachment of natural lightning flashes to trees: preliminary statistical characteristics. *Journal of Lightning Research, 1*, 9-21.

Martinetti, D., & Geniaux, G. (2017). Approximate likelihood estimation of spatial probit models. *Regional Science and Urban Economics, 64*, 30-45.

Martínez, J., Martínez-Vega, J., & Martín, M. P. (2004). El factor humano en los incendios forestales: Análisis de factores socio-económicos relacionados con la incidencia de incendios forestales en España. In E. Chuvieco & P. Martín (Eds.), *Nuevas tecnologías para la estimación del riesgo de incendios forestales* (pp. 101–142). Madrid (Spain): Colección de Estudios Ambientales, CSIC.

Martínez-Fernández, J., Chuvieco, E., & Koutsias, N. (2012). Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. *Natural Hazards and Earth System Sciences, 12,* 1–17.

McGuiney, E., Shulski, M., & Wendler, G. (2005). Alaska lightning climatology and application to wildfire science. In *Proceedings of Conference on Meteorological Applications of Lightning Data*. San Diego, CA.

McMillen, D. P. (1992). Probit with spatial autocorrelation. *Journal of Regional Science, 32,* 335-348.

Ministerio de Medio Ambiente. (2003). *Mapa forestal de España Escala 1:50,000 (MFE50) de la Provincia de León*. Madrid (Spain): Organismo Autónomo Parques Nacionales.

Molina-Terrén, D., Xanthopoulos, G., Diakakis, M., Ribeiro, L., Caballero, D., Delogu, G. M., Viegas, D. X., Silva, C. A., & Cardil, A. (2019). Analysis of forest fire fatalities in Southern Europe: Spain, Portugal, Greece and Sardinia (Italy). *International Journal of Wildland Fire*, *28*, 85–98.

Moreira, F., Viedma O., Arianoutsou, M., Curt, T., Koutsias, N., Rigolot, E., Barbati, A., Corona, P., Vaz, P., Xanthopoulos, G., Mouillot, F., & Bilgili, E. (2011). Landscape e wildfire interactions in southern Europe: Implications for landscape management. *Journal of Environmental Management*, *92*, 2389-2402.

Müller, M. M., Vacik, H., Diendorfer, G., Arpaci, A., Formayer, H., & Gossow, H. (2013). Analysis of lightning-induced forest fires in Austria. *Theoretical and Applied Climatology, 111,* 183-193.

Mundo, I. A., Wiegandc, T., Kanagarajc, R., & Kitzbergerd, T. (2013) Environmental drivers and spatial dependency in wildfire ignition patterns of northwestern Patagonia. *Journal of Environmental Management, 123*, 77–87.

Nafría, D. A., Garrido, N., Álvarez, M. V., Cubero, D., Fernández, M., Villarino, I., Gutiérrez, A., & Abia I. (2013). *Atlas Agroclimático Castilla y León*. Madrid (Spain): Agencia Estatal de Meteorología (AEMET) and Instituto Tecnológico Agrario de Castilla y León (ITACyL).

Nagh, A., Murphy, M. J., Schulz, W., & Cummins, K. L. (2005). Lightning locating systems: Insights on characteristics and validation techniques. *Earth and Space Science, 2,* 65–93.

Nieto, H., Aguado, I., García, M., & Chuvieco, E. (2012). Lightning-caused fires in Central Spain: Development of a probability model of occurrence for two Spanish regions. *Agricultural and Forest Meteorology, 162*, 35-43.

Oliveira, S., Oehler, F., San Miguel-Ayanz, J., Camia, A., & Pereira, J. M. C. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *Forest Ecology and Management*, *275*, 117–129.

Pace, R. K., & Barry, R. (1997). Quick computation of regressions with a spatially autoregressive dependent variable. *Geographical Analysis, 29*, 232-237.

Pacheco, C. E., Aguado, I., & Nieto, H. (2009). Análisis de ocurrencia de incendios forestales causados por rayo en la España peninsular. *Geofocus, 9*, 232-249.

Pérez-Puebla, F. (2004). Cooperación entre las redes de rayos de España y Portugal. In *Proceedings of Jornadas Científicas de la Asociación Meteorológica Española*. Badajoz (Spain).

Pezzatti, G. B., Bajocco, S., Torriani, D., & Conedera, M. (2009). Selective burning of forest vegetation in Canton Ticino (southern Switzerland). *Plant Biosystems, 143,* 609–620.

Pineda, N., Montanyà, J., & van der Velde, O. A. (2014). Characteristics of lightning related to wildfire ignitions in Catalonia. *Atmospheric Research, 135–136,* 380–387.

Plucinski, M. P. (2013). Modeling the probability of Australian grassfires escaping initial attack to aid deployment decisions. *International Journal of Wildland Fire, 22,* 459–468.

R Development Core Team. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, N., Duff, T. J., & Taylor, P. G. (2018). A lightning-caused wildfire ignition forecasting model for operational use. *Agriculture and Forest Meteorology, 253-254*, 233-246.

Reineking, B., Weibel, P., Conedera, P., & Bugmann, H. (2010). Environmental determinants of lightning- v. human-induced forest fire ignitions differ in a temperate mountain region of Switzerland. *International Journal of Wildland Fire, 19,* 541–557.

Rigolot, E., Fernandes, P., & Rego, F. (2009). Managing wildfire risk: prevention, suppression. In Y. Birot (Ed.), *Living with Wildfire: What Science Can Tell Us* (pp. 59-52). European Forest Institute.

Rivas-Soriano, L., de Pablo, F., & Tomas, C. (2005). Ten-year study of cloud-to-ground lightning activity in the Iberian Peninsula. *Journal of Atmospheric and Solar-Terrestrial Physics, 67,* 1632–1639.

Robla-González, E., Vallejo-Bombín, R., De La Cita-Benito, F. J., & Lerner-Cuzzi, M. (2009). El mapa forestal de España a escala 1:50.000 (1998-2007): Resumen y resultados de un proyecto. In SECF & Junta de Castilla y León (Eds.), *Proceedings of 5º Congreso Forestal Español*. Ávila (Spain).

Rorig, M. L., & Ferguson, S. A. (1999). Characteristics of lightning and wildland fire ignition in the Pacific Northwest. *Journal of Applied Meteorology, 38,* 1565–1575.

San Miguel-Ayanz, J., Carlson, J. D., Alexander, M., Tolhurst, K., Morgan, G., Sneeuwjagt, R., & Dudley, M. (2003). Current methods to assess fire danger potential. In E. Chuvieco (Ed.),

Wildland fire danger estimation and mapping. The role of remote sensing data (pp. 21-61). Singapore, Republic of Singapore: World Scientific Publishing.

Satir, O., Berberoglu, S., & Donmez, C. (2016). Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomatics, Natural Hazards and Risk, 7*, 1645-1658.

Smith, T. E. (2016). *Notebook on spatial data analysis*. Retrieved from https://www.seas.upenn.edu/~ese502/.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostics systems: methods from signal detection theory*. New York: Academic Press.

Vacik, H., & Müller, M. M. (2017). Characteristics of lightnings igniting forest fires in Austria. *Agricultural and Forest Meteorology, 240,* 26-34.

Van Rijsbergen, C. J. (1979). *Information Retrieval (2nd ed.)*. London, UK: Butterworth-Heinemann.

Vankat, J. L. (1985). General patterns of lightning ignitions in Sequoia National Park, California. In *Proceedings of Symposium and Workshop on Wilderness Fire, USDA Forest Service General Technical Report INT-182* (pp. 408–411). Missoula, MT.

Vasconcelos, M. J. P., Silva, S., Tomé, M., Alvim, M., & Pereira, J. M. (2001). Spatial prediction of fire ignition probabilities: comparing logistic regression and neural networks. *Photogrammetric Engineering and Remote Sensing, 67,* 73–83.

Vázquez, A., & Moreno, J. M. (1998). Patterns of lightning- and people-caused fires in Peninsular Spain. *International Journal of Wildland Fire*, *8*, 103–115.

Vecín-Arias, D., Castedo-Dorado, F., Ordóñez, C., & Rodríguez-Pérez, J. R. (2016). Biophysical and lightning characteristics drive lightning-induced fire occurrence in the central plateau of the Iberian Peninsula. *Agricultural and Forest Meteorology, 225*, 36-47.

Vega-García, C., & Chuvieco, E. (2006). Applying local measures of spatial heterogeneity to Landsat-TM images for predicting wildfire occurrence in Mediterranean landscapes. *Landscape Ecology, 21*, 595–605.

Vilar, L., Woolford, D. G., Martell, D. L., & Martín, P. M. (2010). A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. *International Journal of Wildland Fire, 19,* 325–337.

Wilhelm, S., & Godinho de Matos, M. (2013). Estimating Spatial Probit Models in R. *The R Journal, 5,* 130–143.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science.

Wood, S. N., & Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling, 157,* 157-177.

Woolford, D. G., Bellhouse, D. R., Braun, W. J., Dean, C. B., Martell, D. L., & Sun, J. A. (2011). Spatio-temporal model for people-caused forest fire occurrence in the Romeo Malette Forest. *Journal of Environmental Statistics, 1*, 1-26.

Yanoviak, S. P., Gora, E. M., Fredley, J., Bitzer, P. M., Muzika, R. M., & Carson, W. P. (2015). Direct effects of lightning in temperate forests: a review and preliminary survey in a hemlock–hardwood forest of the northern United States. *Canadian Journal of Forest Research, 45,* 1258–1268.

Tables

Table I. Variables computed for each 4×4-km grid cell in the studied area

Variable	Depiction
%coniferous	% of coniferous woodlands
%broadleaf	% of broadleaf woodlands
%mixed	% of mixed forests
%shrub	% of shrublands
%crop	% of agricultural croplands
%pasture	% of pasturelands
%other	% of non-combustible areas
altitude	mean altitude (m)
slope	mean slope (%)
%north	% with north-facing aspect
%east	% with east-facing aspect
%south	% with south-facing aspect
%west	% with west-facing aspect
%flat	% with flat terrain
total flashes	Averaged annual number of flashes (flashes km^{-2})
flashes -	Averaged annual number of flashes with negative polarity (flashes km^{-2})
flashes +	Averaged annual number of flashes with positive polarity (flashes km^{-2})
peak current -	Averaged annual mean peak current of flashes with negative polarity (kA)
peak current +	Averaged annual mean peak current of flashes with positive polarity (kA)
thunderstorm days	Averaged annual number of thunderstorm days
dry thunderstorm days	Averaged annual number of dry thunderstorms days

Table II. Prediction performance for the combinations of methods and covariates (q =1 to 6). The covariates included in each combination are indicated in grey.

#covars	Model	Rank within model	%coniferous	%crops	%shrub	%mixed	%North	%other	%pasture	altitude	S_1	S_2	$F1$ score	*Sens*	*Spec*	*CCR*
q = 1	GLM	28	■										0.316	0.452	0.861	0.824
	RFM	35	■										0.234	0.381	0.815	0.776
	GAM	28	■										0.316	0.452	0.861	0.824
	AUREG	26	■										0.336	0.524	0.842	0.814
	GAMs	32	■										0.371	0.466	0.890	0.851
q = 2	GLM	1	■								■		0.346	0.417	0.902	0.858
	RFM	3									■	■	0.337	0.446	0.881	0.842
	GAM	28	■	■									0.316	0.452	0.861	0.824
	AUREG	1	■	■									0.359	0.482	0.881	0.845
	GAMs	23	■				■						0.364	0.529	0.864	0.834
q = 3	GLM	2	■	■							■		0.342	0.406	0.904	0.857
	RFM	2	■								■	■	0.352	0.452	0.889	0.812
	GAM	1	■	■							■		0.343	0.500	0.860	0.828
	AUREG	3	■	■	■								0.335	0.464	0.886	0.849
	GAMs	7	■	■			■						0.372	0.548	0.862	0.834
q = 4	GLM	4	■	■							■	■	0.339	0.440	0.886	0.846
	RFM	5	■			■					■	■	0.334	0.327	0.938	0.883
	GAM	2	■	■	■						■		0.354	0.559	0.842	0.816
	AUREG	2	■	■	■				■				0.356	0.512	0.866	0.834
	GAMs	4	■			■	■	■					0.381	0.536	0.874	0.843
q = 5	GLM	8	■	■	■	■					■		0.325	0.458	0.865	0.829
	RFM	1	■	■	■						■	■	0.356	0.577	0.835	0.812
	GAM	8	■		■	■					■	■	0.345	0.583	0.822	0.801
	AUREG	8	■	■	■				■	■			0.346	0.464	0.879	0.842

#covars	Model	Rank within model	%coniferous	%crops	%shrub	%mixed	%North	%other	%pasture	altitude	s_1	s_2	*F*1 score	*Sens*	*Spec*	*CCR*
	GAMs	1	■	■		■	■	■					0.406	0.526	0.879	0.843
***q* = 6**	**GLM**	1	■	■	■	■					■	■	0.314	0.506	0.830	0.801
	RFM	7	■	■	■	■					■	■	0.333	0.506	0.849	0.818
	GAM	5	■	■	■	■					■	■	0.347	0.506	0.861	0.829
	AUREG	17	■	■	■	■			■	■			0.334	0.440	0.882	0.842
	GAMs	2	■	■	■	■	■	■					0.345	0.578	0.843	0.821

Note: q is number of predictors in the model; s_1 and s_2 are planar UTM coordinates; *Sens*, *Spec* and *CCR* are the sensitivity, specificity and correct classification rate statistics, respectively. The rank within the model shows the order of $F1$ scores within the method used in model fitting. GAMs models include a bivariate function of planar coordinates s_1 and s_2 as covariates, so strictly speaking q equals q+2 in this table. AUREG models do not include s_1 and s_2 planar coordinates, since they implicitly collect the relative spatial position of observations by means of a weight matrix.

Table III. Estimated linear coefficients and their significance values and estimated significance of the smooth terms for the selected GAMs (q = 1; %coniferous).

Variable	**Estimated parameter**	**Standard error**	***z*-value**	***p*-value**
Intercept	-2.859	0.0845	-33.82	$<2 \times 10^{-16}$
	Estimated degrees of freedom	**Estimated residual degrees of freedom**	**chi-square**	***p*-value**
Bivariate effect (longitude, latitude)	26.96	28.69	145.1	$<2 \times 10^{-16}$
%coniferous	6.424	7.551	103.1	$<2 \times 10^{-16}$

dof: degrees of freedom.

Figure captions

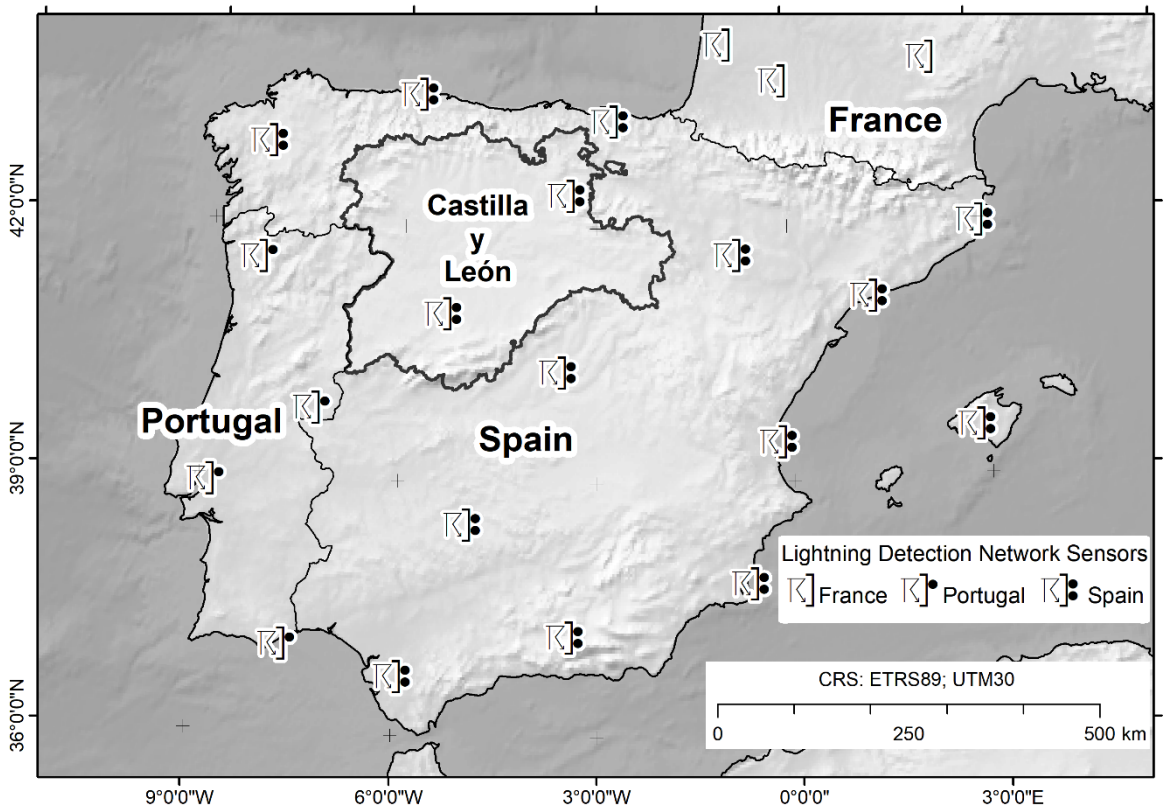


Fig. 1. Location of the study area (Castilla y León) in Spain and of lightning detection network sensors in Spain, Portugal and southern France.

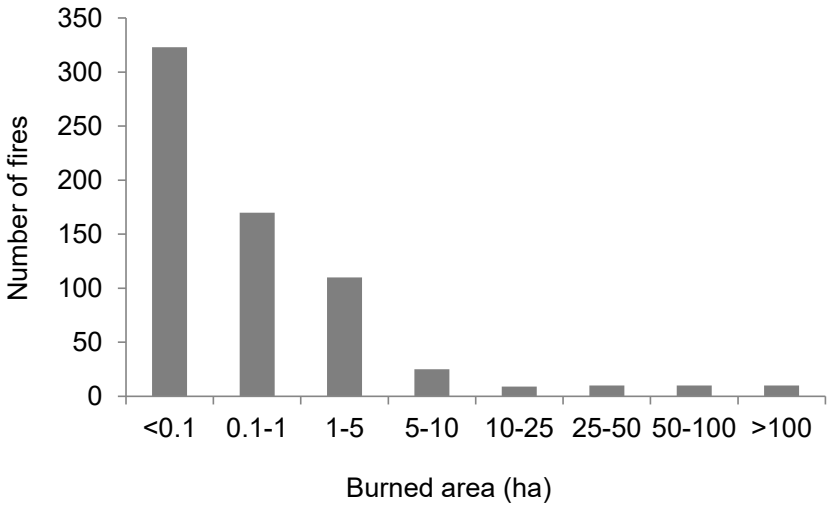


Fig. 2. Distribution by size class of 662 lightning-caused forest fires with available planar coordinates reported for Castilla y León (Spain) in the period 2000-2010.

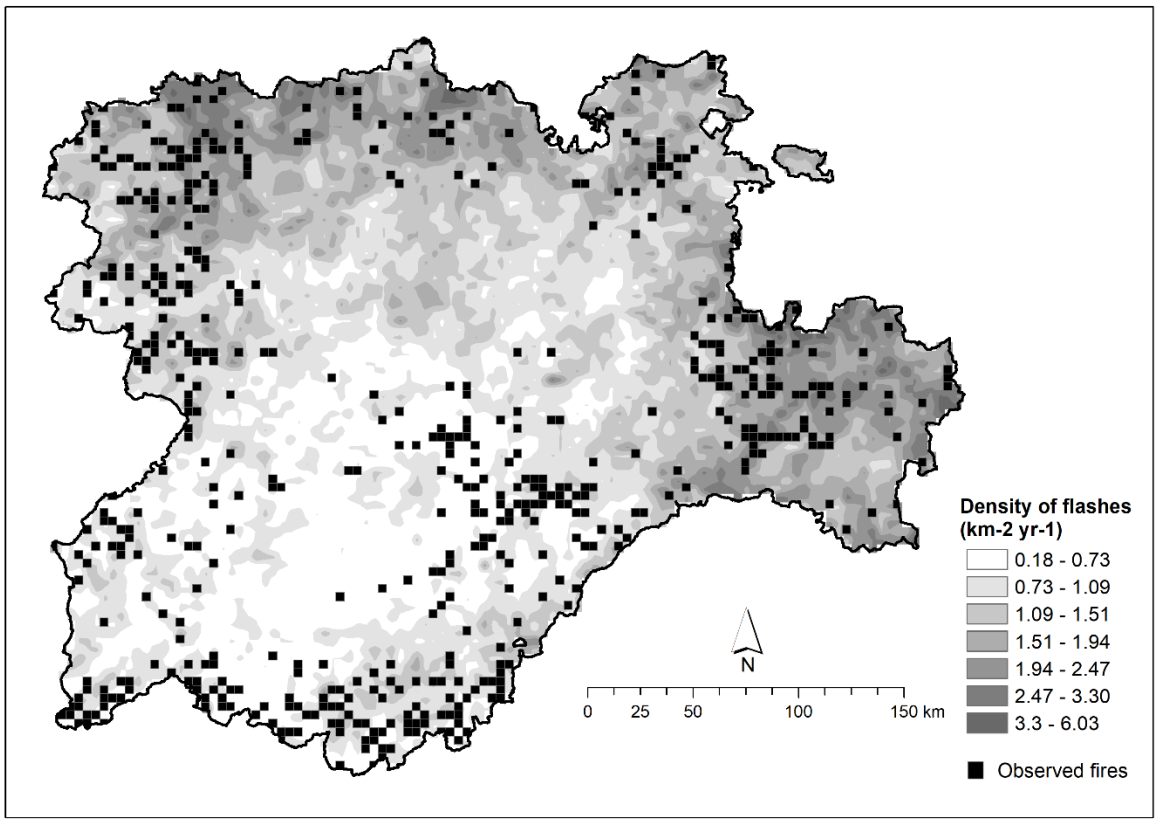


Fig. 3. Spatial distribution in a 4 × 4-km grid of 662 lightning-caused forest fires with available planar coordinates reported for Castilla y León (Spain) in the period 2000-2010 superimposed on the spatial distribution of flash density.

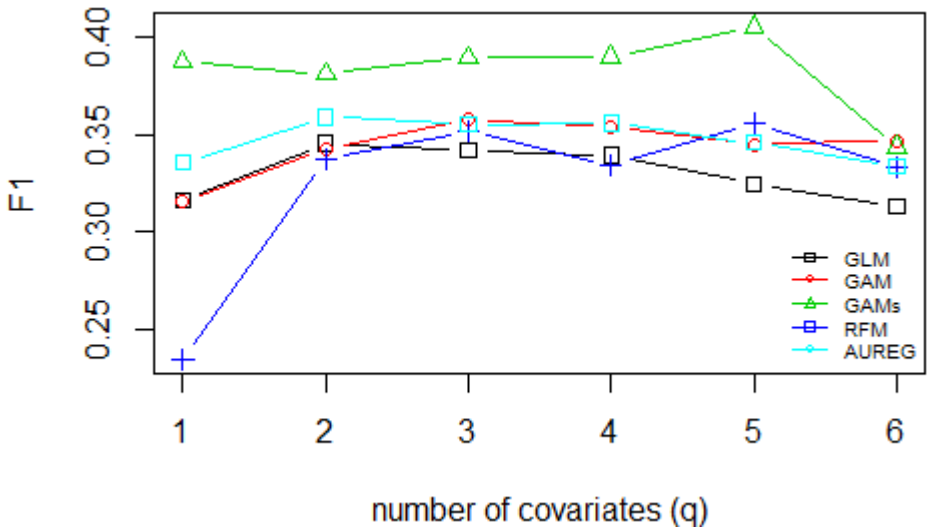


Fig. 4. *F*1 scores for the 5 tested models according to the number of covariates included (q = 1 to 6). The results correspond to the test sample. Note that GAMs models include a bivariate function for the s_1 and s_2 planar coordinates so, strictly speaking, q equals q+2 in this figure.

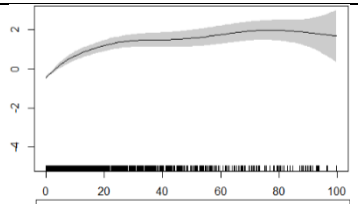
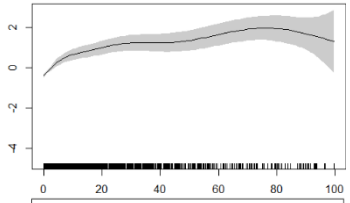
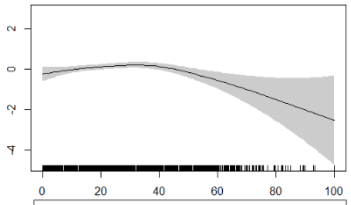
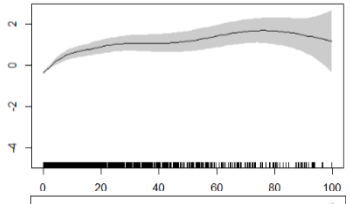
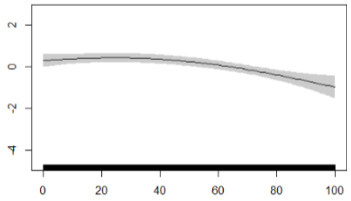
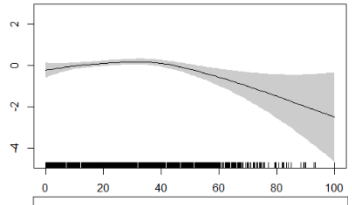
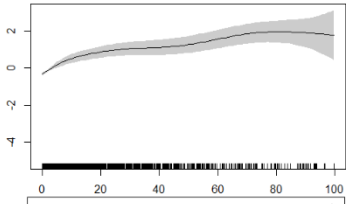
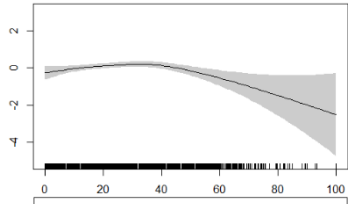
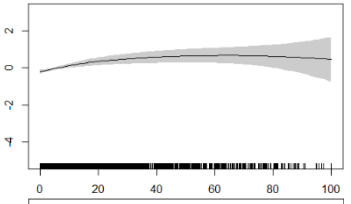
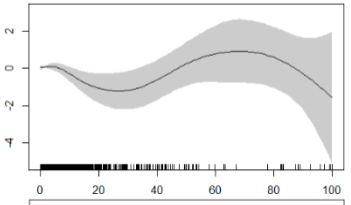
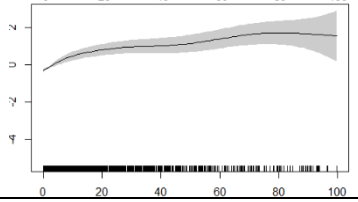
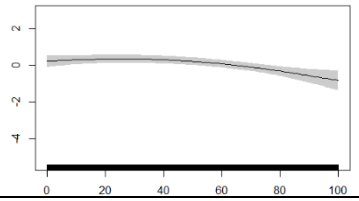
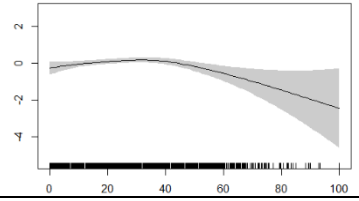
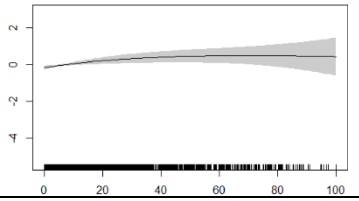
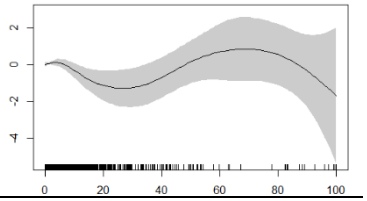
#covars	%coniferous	%crops	%North	%mixed	%other
$q = 1$		-	-	-	-
$q = 2$		-		-	-
$q = 3$				-	-
$q = 4$		-			
$q = 5$					

Fig. 5. Plots for estimated partial dependence (solid black lines) along with 95% confidence bands (grey areas) for covariates included in the GAMs models for q = 1 to 5. The horizontal axis represents the values of the explanatory variable and the vertical axis represents the values of $logit = \left\{log\left(\frac{P}{1-P}\right)\right\}$, where P represents the probability of a lightning-caused fire. Plots were generated by running the GAMs models for the entire dataset.

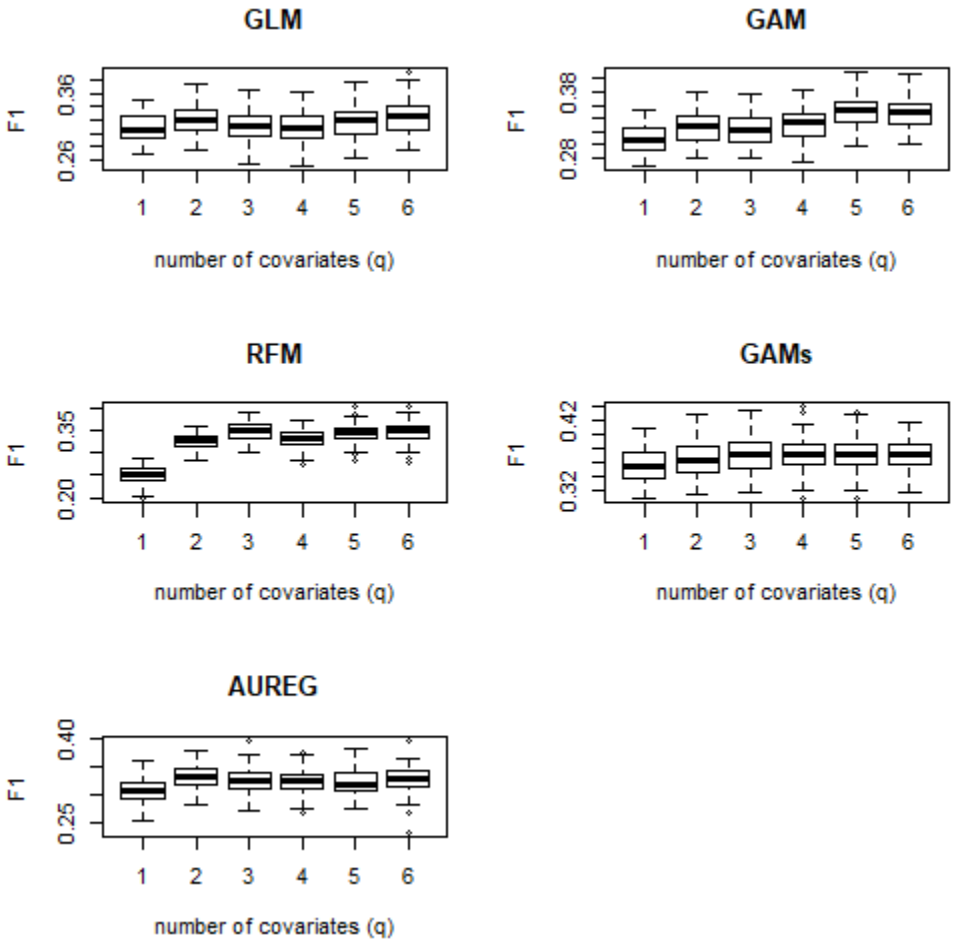


Fig. 6. Boxplots showing the distribution of *F*1 scores for each of the five models according to the number of covariates (q = 1 to 6). Note that GAMs models include a bivariate function for the s_1 and s_2 planar coordinates so, strictly speaking, q equals q+2 in this figure.

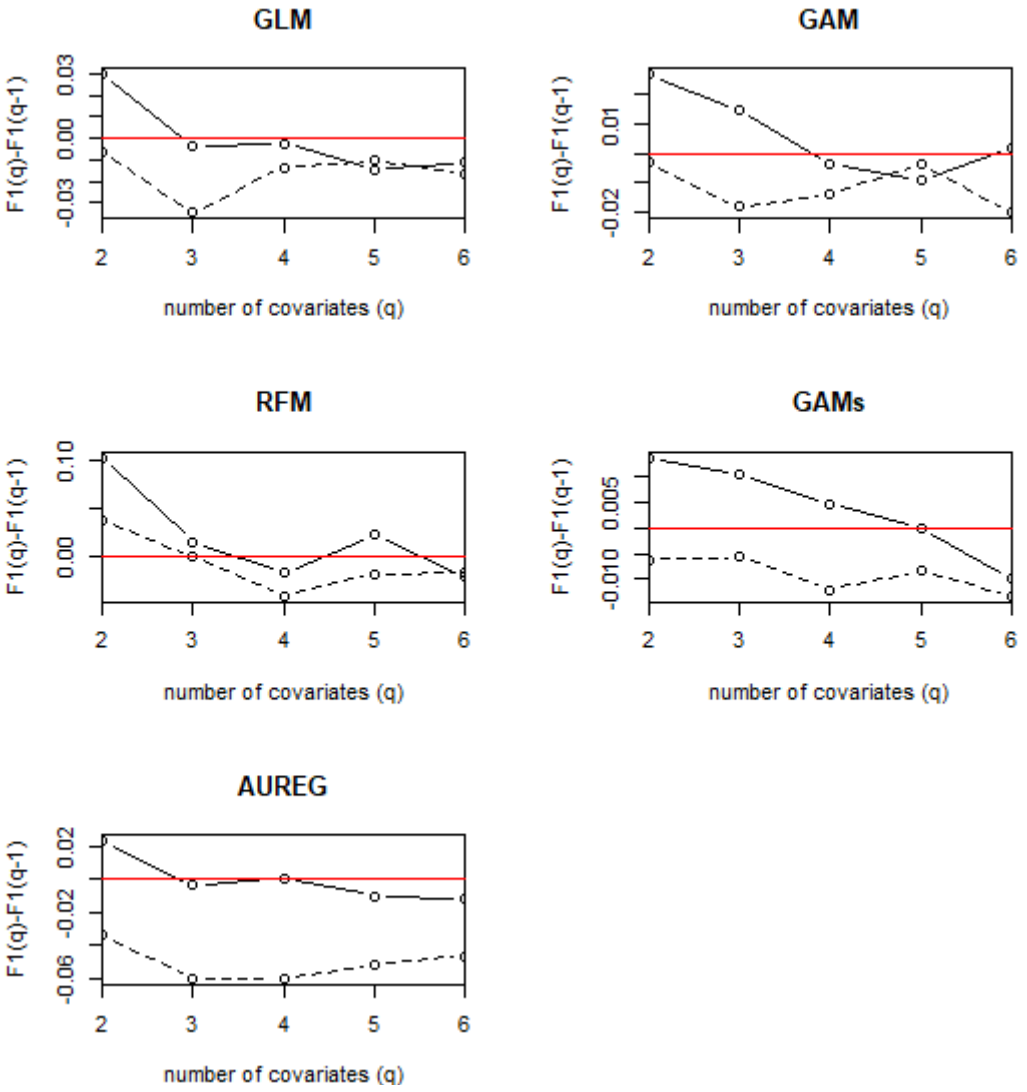


Fig. 7. $F1$ score differences, $F1(q) - F1(q\text{-}1)$, for the five models according to the number of covariates q (solid lines) and lower limit of the confidence interval $[a,\infty]$ for the differences (dashed lines). The horizontal red lines mark the zero value. The level of significance used was 0.05. Note that GAMs models include a bivariate function for the s_1 and s_2 planar coordinates so, strictly speaking, q equals q+2 in this figure.

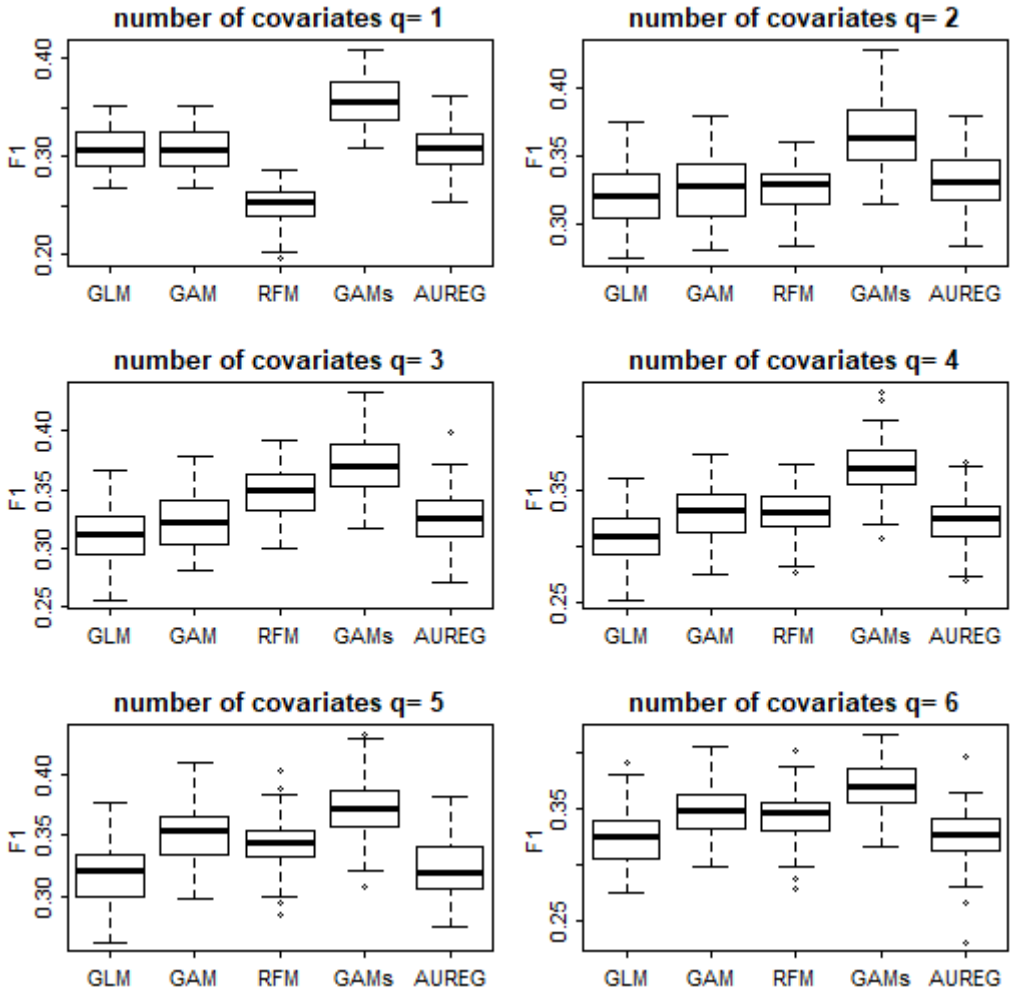


Fig. 8. Boxplots showing the distribution of *F*1 scores for the five models fixing the number of covariates (q = 1 to 6). Note that GAMs models include a bivariate function for the s_1 and s_2 planar coordinates so, strictly speaking, q equals q+2 in this figure.

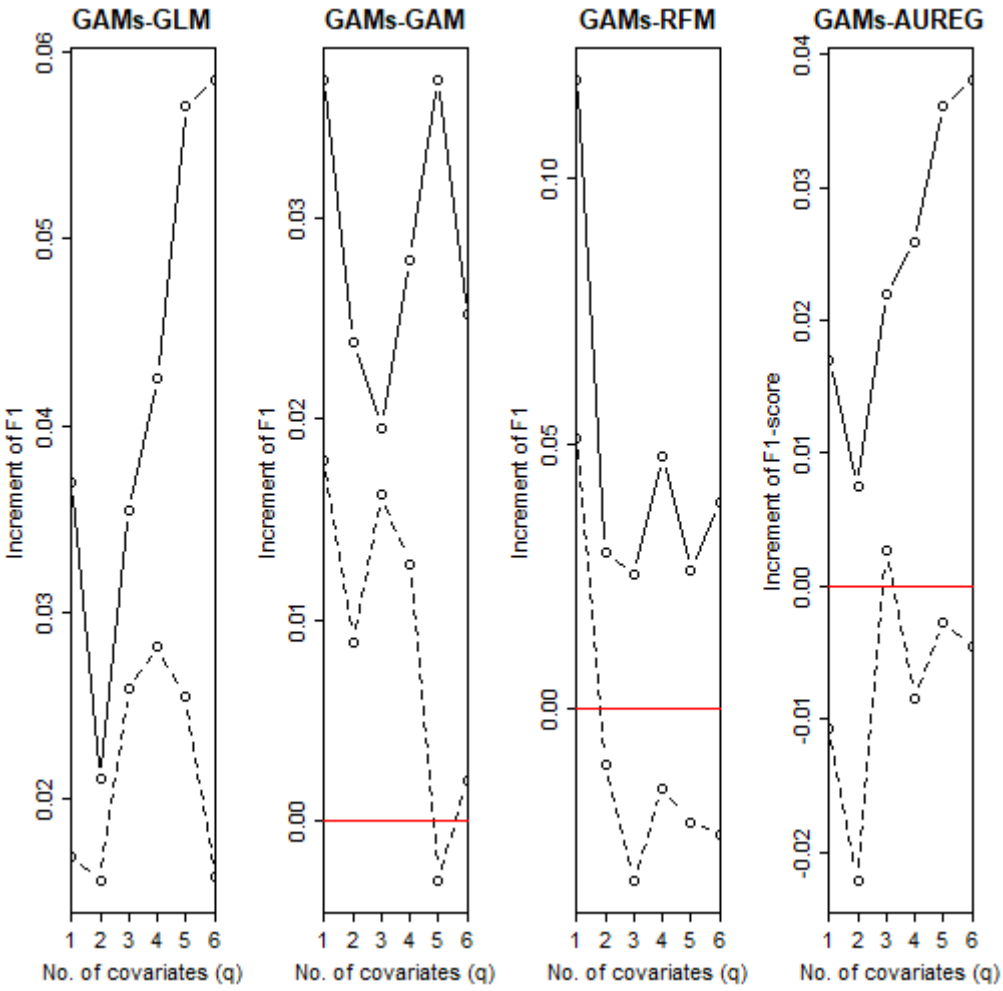


Fig. 9. Increment in $F1$ score when the GAMs is compared with the GLM, GAM, RFM and AUREG methods. The dashed lines represent the lower limit of the confidence interval $[a,\infty]$. The level of significance used was 0.05. Note that GAMs models include a bivariate function for the s_1 and s_2 planar coordinates so, strictly speaking, q equals q+2 in this figure.

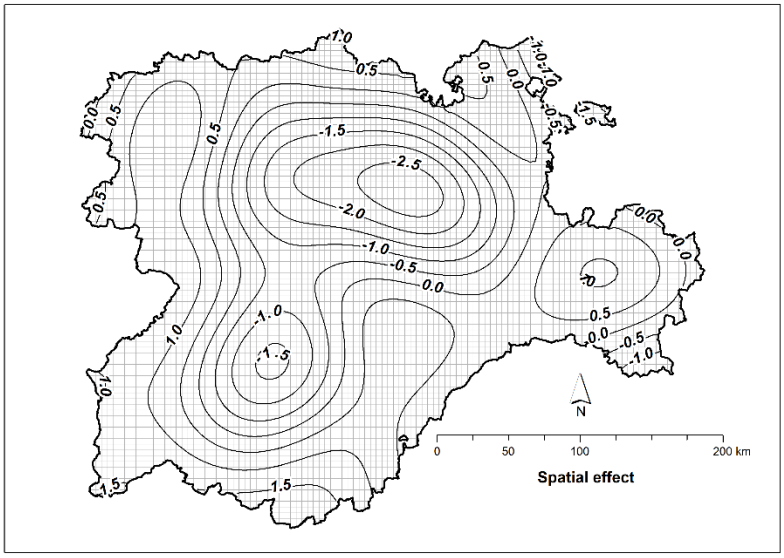


Fig. 10. Estimated spatial effect in contour form generated by running the selected GAMs (q = 1; %coniferous) for the entire dataset.

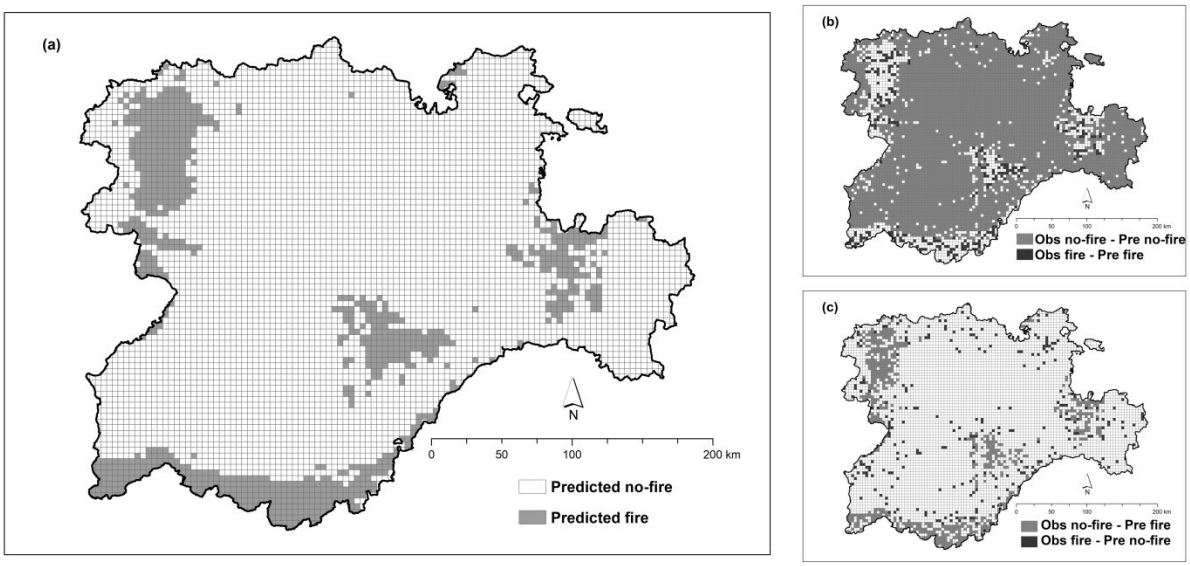


Fig. 11. (a) Predicted spatial distribution of lightning-caused fire occurrence according to the selected GAMs (q = 1; %coniferous). **(b)** Correctly classified grid cells of lightning-caused fire occurrence. **(c)** Incorrectly classified grid cells of lightning-caused fire occurrence. All the figures were generated by running the selected GAMs for the entire dataset.

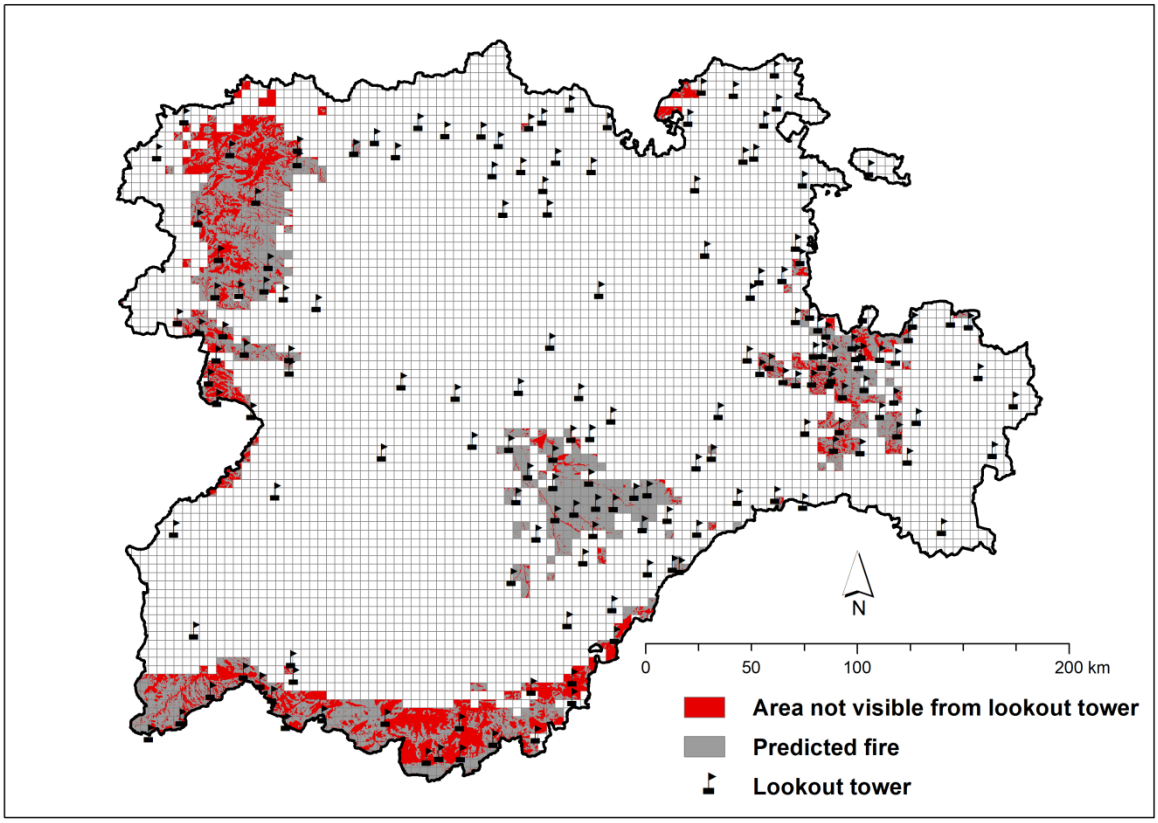


Fig. 12. Visibility analysis map of the current lookout tower network superimposed on the predicted spatial distribution of lightning-caused fire occurrence according to the selected GAMs ($q = 1$; %coniferous).