Facultad de Ciencias

Automated wildfire season detection at a global scale: Application for the development of a predictive system of fire activity

Detección automática de la estación de fuegos a escala global: Aplicación para el desarrollo de un sistema predictivo de incendios

Trabajo Fin de Máster
para acceder al
Máster Universitario en Ciencia de Datos

Presentado por
Marcos Valle Miñón
Dirigido por
Joaquín Bedia Jiménez
Rodrigo García Manzanas

Septiembre 2021

Resumen

En la primera parte de este trabajo describimos un procedimiento de aprendizaje automático no supervisado basado en la técnica de *Gaussian Mixtures* con el objetivo de determinar la estación de fuegos a escala global a partir de datos de satélite de área quemada a una resolución espacial de 0.5°. Nuestro método permite la identificación de ciclos anuales de tipo unimodal y multimodal, así como la determinación del inicio, fin y momento de máxima actividad de incendios, con la ventaja adicional de proporcionar un procedimiento totalmente automatizado que puede ser utilizado a múltiples escalas espacio-temporales. La caracterización de la estación de fuegos aquí presentada desvela un patrón inequívoco y espacialmente coherente, que es consistente con estudios previos sobre la estacionalidad de incendios. Nuestro método puede ser fácilmente adaptado por el usuario mediante el ajuste de unos pocos parámetros sencillos para adecuarlo a bases de datos de incendios de distinta naturaleza, extensión geográfica y resolución espacio-temporal.

A continuación, se parte de la zonificación proporcionada por las *Gaussian mixtures* para el desarrollo de modelos predictivos de área quemada —durante la estación de fuegos principal— a nivel de clúster, tomando como única información predictora una serie de índices que representan los patrones de teleconexión climática más relevantes a escala global. Para ello se consideran modelos lineales, *random forest* y k-vecinos cercanos, en cuyo ajuste se aplican técnicas de validación cruzada. Nuestros resultados muestran que, cuando se consideran como predictores aquellos índices que están más fuertemente correlacionados con el área quemada, incluso los modelos lineales más simples son capaces de proporcionar predicciones de área quemada fiables en determinadas zonas del planeta. Este trabajo abre la puerta para el futuro desarrollo e implementación de un sistema operativo de alerta temprana de incendios en base a modelos climáticos de predicción estacional.

Todos los análisis realizados son totalmente reproducibles a través de los datos post-procesados, scripts y notebooks que están disponibles en un repositorio público.

Palabras clave: *área quemada, estación de incendios, clustering, patrones de teleconexión, modelos climáticos empíricos de incendios, minería de datos*

Abstract

In the first part of this project, we describe an unsupervised machine learning procedure based on *Gaussian mixtures* in order to determine the fire season at a global scale, using remotely sensed data of burned area at a 0.5° spatial resolution. Our results allow the identification of unimodal and multimodal annual cycles as well as the start, end and timing of bulk fire activity, with the added advantage of providing a fully automated procedure that can be used at multiple temporal and spatial scales. The fire season characterization presented unveils an unambiguous, spatially coherent pattern, consistent with previous studies on fire seasonality. Our method can be easily tuned by the user through the manipulation of a few simple parameters in order to accommodate fire databases of varying nature, geographical extent and spatial and temporal resolutions.

Next, using the fire season definition given by the *Gaussian mixtures*, predictive models of burned area are developed at a global scale using a set of the most relevant climate teleconection indices as predictors. We consider linear models, random forests and k-nearest neighbours, fitted following a cross-validation setup. Our results show that, when only the most correlated indices with the fire intensity are considered as predictors, even the simplest linear models are able to give accurate predictions in certain parts of the world. This study paves the way for the implementation of an operational early-warning wildfire system based on seasonal forecasting climate models.

All the analyses undertaken are fully reproducible through the post-processed data, scripts and notebooks available through a dedicated open repository.

Keywords: *burned area, fire season, clustering, teleconnection patterns, empirical fire-climate models, data mining*

Agradecimientos

Quiero aprovechar este trabajo para agradecer a todas las personas que me han ayudado durante estos años en la universidad. En primer lugar, a mis padres y a mi hermana por su apoyo diario. También quiero dar las gracias a mi pareja y a mis amigos más cercanos por estar siempre ahí y a mis compañeros del máster, que siempre me han echado una mano cuando me hacía falta.

Además, quiero agradecer todo su trabajo y dedicación a mis dos directores.

Por último, quiero dedicar este trabajo a mi abuela, que es para mí un ejemplo a seguir.

Contents

CHAPTER 1

Introduction

1.1 Fire Seasonality

Fire is a global-scale phenomenon directly affecting different components of the Earth System such as the structure and distribution of vegetation, the composition of the atmosphere, hydrosphere and soils, the global biogeochemical cycles and the climate system (Bond et al., 2004). As such, it is a complex Earth System Process (Bowman et al., 2009) with an heterogeneous spatio-temporal distribution across the globe (Chuvieco et al., 2008) driven by a variety of spatial and environmental gradients (Krawchuk and Moritz, 2010; Pausas and Ribeiro, 2013), including anthropogenic factors acting either indirectly (Bowman et al., 2011) or directly through ignition as part of land management practices (see e.g.: Magi et al., 2012; Pereira et al., 2015).

As a result, global land areas are affected by their own particular *fire regimes*, characterized by the frequency, intensity, seasonality, extent and type of fires at different spatial and temporal scales (Archibald et al., 2013). Describing fire seasonality is therefore crucial for a better understanding of fire regimes across different world regions (Le Page et al., 2010), serving as a tool for the characterization of inter-annual cycles and fire activity peaks (Boschetti and Roy, 2008), and as a baseline for undertaking interannual fire analyses requiring some form of seasonal aggregated statistics (for instance in climate change impact studies, e.g.: Bedia et al., 2015; Jolly et al., 2015), as well as a reference for the assessment of fire models'

ability to reproduce seasonal patterns of burned area (Kelley et al., 2013; Hantson et al., 2020).

With the recent availability of remotely sensed global fire data products of an adequate quality and temporal coverage (Giglio et al., 2013), some previous efforts to characterize the fire season shape globally have been undertaken. Their aim is mostly focused on detecting multimodalities in the annual fire cycle building on parametric statistical tests (Ameijeiras-Alonso et al., 2019), signaled as a sign of the "human footprint" in fire regimes (Le Page et al., 2010; Benali et al., 2017). In this work, a novel method based on Gaussian mixtures is presented in order to automatically compute the fire season at the pixel-scale. The main underlying assumption is that the annual fire cycle can be reproduced through a mixture of Gaussian distributions and the method can therefore accommodate unimodal and multimodal responses without supervision, given a sample of sufficient size and quality as the 20-year monthly time series database used here (Sec. 2.1).

1.2 Empirical Fire-climate Models

In the framework of wildfire prevention, even marginal improvements in suppression efficiency have the potential to prevent significant damages and economic costs derived from wildfires (Preisler and Westerling, 2007). Therefore, seasonal predictions have a great potential to aid decision-making (see e.g. Bedia et al., 2018; Turco et al., 2018), helping fire agencies to improve the efficiency of wildfire suppression efforts during severe fire seasons and optimize the available economic, technical and human resources through the provision of actionable information. Some of these models rely on specific fire danger indices or lagged meteorological variables used as predictors for burned area (see e.g. Bedia et al., 2014; Marcos et al., 2015).

At a global scale, large-scale sea-surface temperature (SST) patterns have been identified as drivers of fire activity over vast land areas (Chen et al., 2016). In this regard, the so called "climate teleconnections" are prominent modes of variability, often linked to SST and/or sea level pressure (SLP) variability in certain regions that exert an influence on weather conditions in distant parts of the world (Barnston and Livezey, 1987). These teleconnections are unveiled by significant statistical links of specific climate indices with meteorological variables (precipitation, temperature, ...) in the region of interest (see Sec. 2.3 for further detail). For this reason, teleconnection indices can be used to establish empirical links with weather-dependent phenomena such as fire activity, providing in this case a potential tool for anticipating fire activity using lagged statistical models (Rodrigues et al., 2021).

Despite the potential of teleconnection patterns for fire activity forecasting on seasonal time scales, to date we are not aware of any previous studies addressing the predictability of fire activity relying solely on climate teleconnection indices globally (however, see Rodrigues et al., 2021, for an application on the Iberian Peninsula). One major limitation for this task is the characterization of a local fire season, in order to focus the analyses on the period of the year of relevance for each location of the world. In this work, we develop such empirical models exploiting the results from the first part of the study that provide, for each global land area pixel, a precise definition of the fire season.

1.3 Objectives

Based on the previous considerations, this work poses two fundamental objectives:

- Identify homogeneous regions in terms of their fire regimes across the world using an automated clustering approach, providing a precise fire season definition at the pixel scale.
- Develop a global predictive system for the severity of the oncoming fire season based on climate teleconnection indices.

1.4 Structure

Chapter 2 describes the data used for the elaboration of this work. The clustering procedure and the techniques considered for the development of predictive models for fire activity are described in Chapter 3. The results obtained are shown and discussed throughout Chapter 4. Finally, the main conclusions are outlined in Chapter 5, together with some future research lines and research reproducibility information.

CHAPTER 2

Data

2.1 Global Fire Database

We have used monthly data of Burned Area (BA) at 0.5° resolution from the *Fire burned area from 2001 to present derived from satellite observations* database (DOI: 10.24381/cds.f333cf85) which is publicly available through the Copernicus Climate Data Store (`https://cds.climate.copernicus.eu/cdsapp#!/home`) as part of Copernicus, the European Union's Earth Observation Programme managed by the European Commission (`https://www.copernicus.eu/en`).

The BA data used are derived through the analysis of reflectance changes from the medium resolution sensors Terra MODIS and Sentinel-3 OLCI, helped by the use of MODIS thermal information. The algorithms used are adapted to the native data from these sensors to produce an homogeneous gridded dataset of global coverage containing monthly data of BA at the pixel scale, extending the database to the present. A more detailed data description is provided in the database landing page at the Climate Data Store (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-fire-burned-area?tab=overview`).

We have worked with the total monthly BA in hectares, as depicted in Fig. 2.1. The data encompasses the period January 2001 to April 2020. In addition, we downloaded the variable "Fraction of Burnable Area" (*fba*), that was used to mask global land areas with very low fuel cover to be discarded from the analyses. In particular, we masked out all pixels with less than 10% *fba*.

Fire data processing has entailed the download from the Climate Data Store and further post-processing including NetCDF file extraction (a compressed binary data format, see `https://www.unidata.ucar.edu/software/netcdf/`) and conversion to a standard R data structure (a data frame), data collocation (geo-location and date handling), conversion from m^2 to *hectares*, data visualization (Fig. 2.1) and masking, prior to data analysis. These tasks have been undertaken with the climate4R framework for climate and geoscientific data analysis and visualization (Frías et al., 2018; Iturbide et al., 2019).

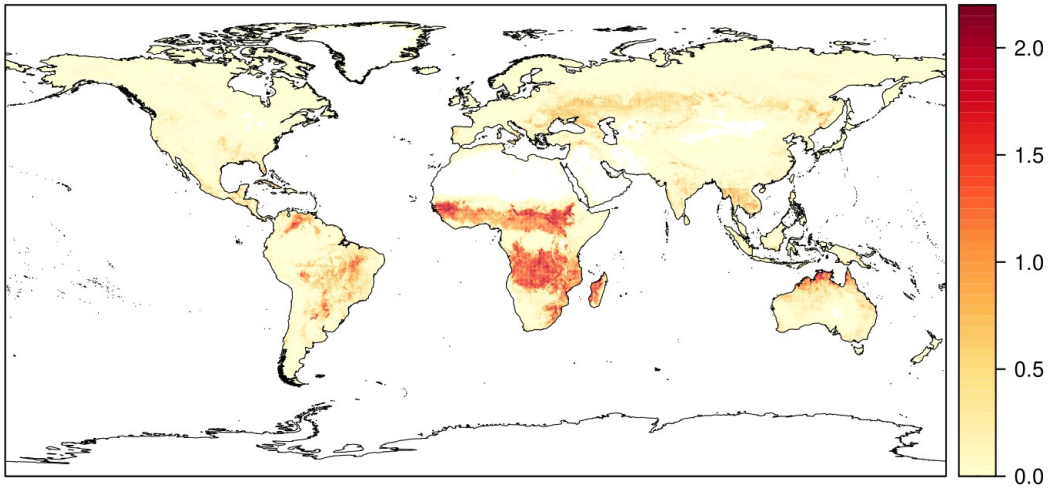

Figure 2.1: Total annual BA (in $\log_{10}$-transformed hectares), averaged for the period of study 2001-2020.

2.2 Biomes

Biomes constitute a large-scale global land surface classification based on the similarity of environmental conditions, including vegetation and bioclimatic characteristics. Therefore, they conform suitable aggregation units for clustering as they implicitly bring into consideration the fuel types and the main fuel-climate relationships. Consequently, some degree of similarity regarding seasonal fire cycles can be expected between areas of the same biome. As such, in order to improve clustering performance, we have divided the global land pixels according to the biomes they belong to, based on the Global Terrestrial Ecoregions delineated by Olson et al. (2001). In particular, after excluding a few regions with low interest for fire activity (ice caps of Antarctic region and Greenland, mangroves etc.), we have retained the 13 global biomes listed in Table 2.2, whose spatial distribution is shown in Fig. 2.2),

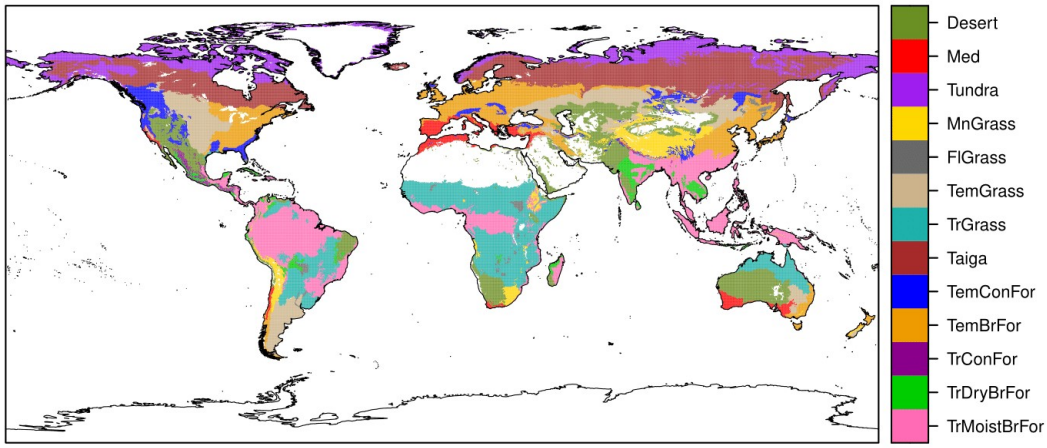

Figure 2.2: Global land biomes (see Table 2.2 for full names and extension). Blank areas correspond to the pixels that were masked out, according to the 10% fraction of burnable area criterion (Sec. 2.1), mostly corresponding to bare rocks, continental water bodies, ice caps and deserts (tropical coastal mangroves are also masked).

2.3 Climate Indices

During the last decades of the past century, the scientific community identified a number synoptic patterns which can promote the development of travelling atmospheric waves which can have important effects in the climate of remote regions with a time-difference of weeks-to-months (these mechanisms are known as climate teleconnections). Many of these patterns have an associated "teleconnection index" which describes their evolution across time. Typically, these teleconnections indices (or simply climate indices) are derived from SLP and SST. For instance, the most popular teleconnection index, El Niño, is based on the SST of the central and eastern tropical Pacific ocean and represents the state of the El Niño-Southern Oscillation (ENSO), the dominant mode of climate variability at seasonal time-scales over the globe (Manzanas et al., 2014).

Building on the potential of these teleconnections to trigger responses in meteorological variables of interest for fire activity —such as temperature, precipitation and winds— throughout the world, this work aims to assess the suitability of a selection of climate indices as explanatory variables for the construction of predictive models for the amount of BA during the fire season, globally (see Sec. 4.2).

The subset of climate indices considered to do so was selected based on the capacity of their corresponding teleconnection patterns to modulate the inter-seasonal and inter-annual variability of climate across many regions of the world. Moreover, monthly anomaly values for all of these indices are publicly available for our pe-

Name	Label	Number of pixels
Tropical and Subtropical Moist Broadleaf Forests	TrMoistBrFor	26429
Tropical and Subtropical Dry Broadleaf Forests	TrDryBrFor	4084
Tropical Conifer Forests	TrConFor	981
Temperate Broadleaf and Mixed Forests	TemBrFor	23411
Temperate Conifer Forests	TemConFor	7550
Boreal Forests/Taiga	Taiga	39775
Tropical and Subtropical Grasslands, Savannas and Shrublands	TrGrass	26344
Temperate Grasslands, Savannas and Shrublands	TemGrass	18459
Flooded Grasslands and Savannas	FlGrass	1540
Montane Grasslands and Shrublands	MnGrass	7714
Tundra	Tundra	28442
Meditarranean Forests, Woodlands and Scrub	Med	5009
Desert and Xeric Shrublands	Desert	21780

Table 2.1: Name, abbreviations and number of non-masked points of the global biomes considered in this study after Olson et al. (2001). Their spatial distribution is depicted in Fig. 2.2.

riod of study (2001-2020) from the Climate Prediction Center (CPC) (`https://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.shtml`) and the National Oceanic and Atmospheric Administration (NOAA) (`https://psl.noaa.gov/data/climateindices/list/`) websites. Note that, working with anomalies instead of absolute values prevents from the appearance of undesired artifacts due to the distinct ranges covered by the different indices. We provide next a brief description of the underlying teleconnection patterns corresponding to the climate indices analyzed in this work, whose monthly time-series along 2001-2020 are shown in Fig. 2.4:

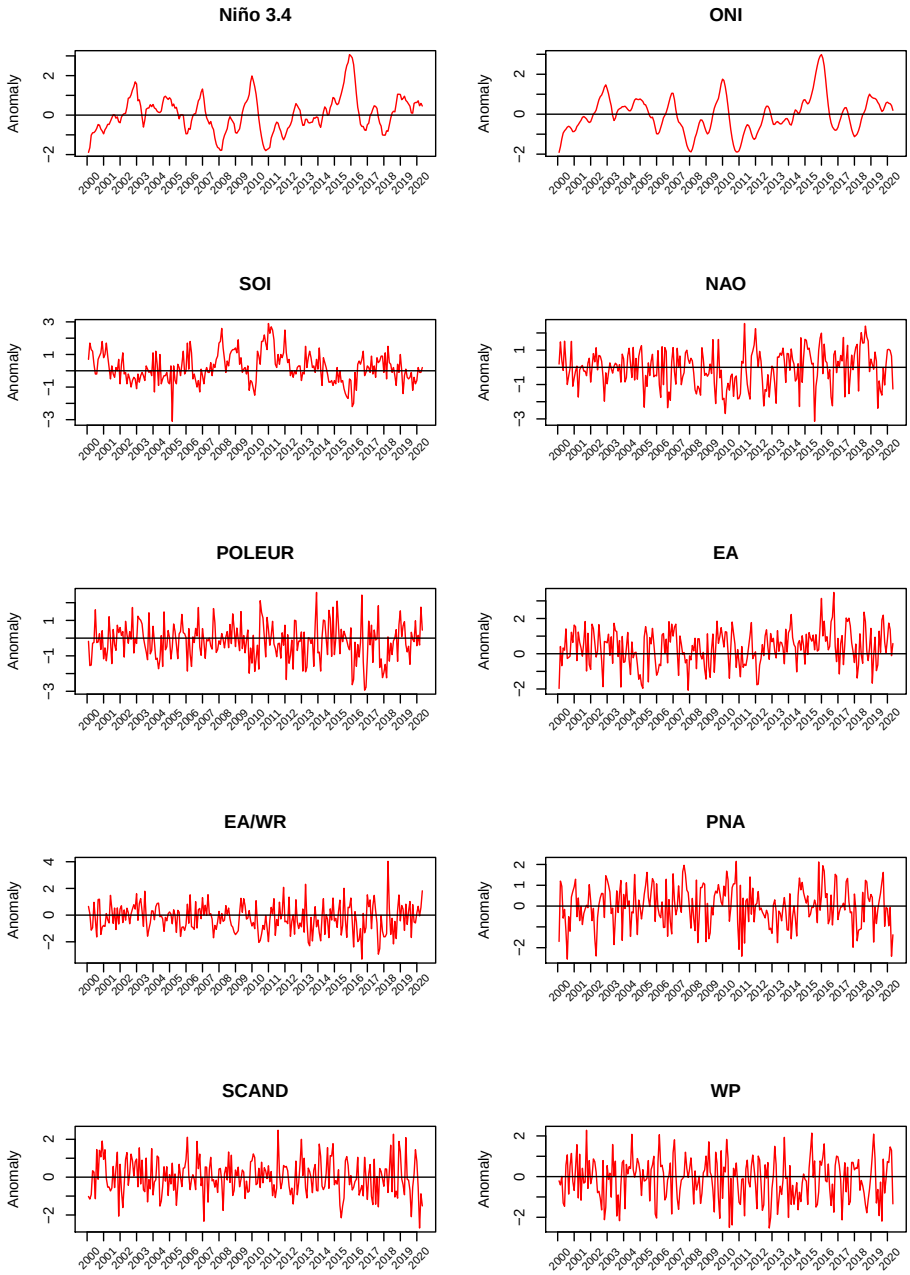

Figure 2.4: Monthly anomalies for all the climate indices considered in this work for the period of study 2001-2020. Note that ONI and Niño 3.4 time-series are smoother as they are based on temperature over the sea-surface, whose thermal inertia is larger than that of land and/or air.

- **Niño 3.4**: This index is defined as the average SST along the East Central Tropical Pacific (5°N-5°S, 120°-170°W), which represents the state of ENSO, a complex climate pattern which brings anomalously warm/cool SSTs over the tropical Pacific recursively every three to seven years. This quasi-periodical

cycle of oscillating SSTs have a strong influence on the distribution of rainfall and temperature not only in the tropics, but in many regions across the globe due to its ability to change the global atmospheric circulation (Taschetto et al., 2020). The warm (cool) phase of ENSO is known as El Niño (La Niña). El Niño (La Niña) conditions usually lead to a slight increase (decrease) in global mean temperature. In addition, during El Niño phase, anomalously dry (wet) conditions are generally found over Australia, northern South America, southern Asia and southern Africa (southwestern North America and eastern Africa). La Niña effects are roughly of opposite sign.

- **Oceanic Niño Index (ONI)**: Together with El Niño 3.4, this is the most commonly used index to monitor El Niño and La Niña phases of ENSO. ONI is defined as the three-month running mean of SST anomalies in El Niño 3.4 region (5°N-5°S, 120°-170°W), based on changing base periods which consist of multiple centered 30-year base periods. The main interest of ONI over the El Niño 3.4 index is that the former is the operational definition used by NOAA, which declares an active El Niño (La Niña) episode when the anomalies exceed 0.5° (-0.5°) for at least five consecutive months.

- **Southern Oscillation Index (SOI)**: This index is computed as the difference between mean SLP anomalies in Tahiti and Darwin, in Australia (Trenberth, 1984). Together with El Niño 3.4 index or ONI, which account for SSTs in the tropical Pacific, the SOI is used to provide a full description of the ENSO phenomenon. Indeed, previous works have demonstrated that both El Niño 3.4 and SOI present a similar quasi-periodicity, with extended periods of negative (positive) SOI corresponding broadly to strong El Nino (La Niña) conditions (Manzanas and Gutiérrez, 2019). Still, the SOI is widely used nowadays to monitor the evolution of ENSO due to its simplicity.

- **North Atlantic Oscillation (NAO)**: This pattern consists of a north-south dipole of SLP anomalies with one center located over Greenland and the other over the North Atlantic (Barnston and Livezey, 1987). The NAO can modulate the intensity and location of the North Atlantic jet stream and storm track (Hurrell, 1995), which in turn results in changes in temperature and precipitation often extending from eastern North America to western and central Europe (Van Loon and Rogers, 1978; Rogers, 1997). In particular, the positive phase of the NAO tends to be associated with above average temperatures in the eastern US and across northern Europe and below average temperatures

in Greenland and oftentimes across southern Europe and the Middle East. It is also related to increased precipitation over northern Europe and Scandinavia in winter and dry conditions over southern and central Europe (Brands et al., 2012). Responses of opposite sign are typically found during its negative phase.

- **Polar/Eurasia (POLEUR)**: This pattern was first described by Esbensen (1984) and formally defined later by Barnston and Livezey (1987). As the NAO, it is also linked to changes in the polar vortex intensity. In particular, its positive (negative) phase consists of negative (positive) SLP anomalies over the polar region (northern China and Mongolia). The POLEUR pattern is associated with above (below) average temperatures in eastern Siberia (eastern China) and enhanced precipitation in the polar region north of Scandinavia.

- **East Atlantic Pattern (EA)**: This pattern was originally defined by Wallace and Gutzler (1981) and later reformulated by Barnston and Livezey (1987). It consists of a north-south dipole of SLP anomalies which is structurally similar to the NAO but shifted to the southeast. The positive phase of EA is associated with above (below) average temperatures in Europe (parts of the US). It is also associated with enhanced (weakened) precipitation over northern Europe and Scandinavia (southern Europe).

- **East Atlantic/ Western Russia (EA/WR)**: As defined by Barnston and Livezey (1987), this pattern consists of four main SLP anomaly centers. Its positive phase —characterized by enhanced SLP over Europe and northern China— is associated with above (below) average temperature over eastern Asia (large portions of western Russia and northeastern Africa). It also tends to produce above (below) average precipitation in eastern China (central Europe). The effects during its negative phase —characterized by weakened SLP in central North Atlantic and north of the Caspian Sea— are in general less pronounced (Kim et al., 2013; Krichak et al., 2014).

- **Pacific North American (PNA)**: This pattern constitutes the most prominent mode of low-frequency variability in the Northern Hemisphere extratropics, affecting particularly the North American continent (Wallace and Gutzler, 1981). It is associated with strong fluctuations in the strength and location of the East Asian jet stream, as well as with variations in the ENSO phenomenon. The PNA pattern consists of a east-west dipole of SLP anomalies located over US. During its positive phase, the cold air residing in Canada is

plunged southeastward, which results in below (above) normal temperatures over the eastern (western) US. The effects on precipitation include the appearance of wetter than normal conditions in the Gulf of Alaska and northwestern US and below average rainfalls across mid and eastern US.

- **Scandinavia (SCAND)**: This pattern consists of a primary circulation center over Scandinavia, with weaker centers of opposite sign over western Europe and eastern Russia/western Mongolia (Barnston and Livezey, 1987; Bueh and Nakamura, 2007). In its positive phase —characterized by increased SLP over Scandinavia and western Russia— anticyclonic activity is suppressed, and below average temperatures are found across central Russia and western Europe. Also, above (below) average precipitation tends to occur across central and southern Europe (Scandinavia) (Zveryaev, 2009).

- **West Pacific (WP)**: This pattern is a primary mode of low-frequency variability over the North Pacific (Barnston and Livezey, 1987; Wallace and Gutzler, 1981). During winter and spring, it consists of a north-south dipole of SLP anomalies with one center located over the Kamchatka Peninsula and the other covering portions of southeastern Asia and the western subtropical North Pacific. In addition, a third anomaly center is present over the eastern North Pacific and southwestern US throughout the year. Similarly to the PNA, the WP pattern modulates the location and intensity of the East Asian jet stream, which in turn can induce important changes in the temperature and precipitation regimes of the North Pacific region. In particular, during its positive phase, increased temperatures are usually found over the lower latitudes of the western North Pacific in winter and spring, and cool conditions affect Siberia in all seasons. With regards to precipitation, wetter (drier) than normal conditions are found over the northern (central) North Pacific.

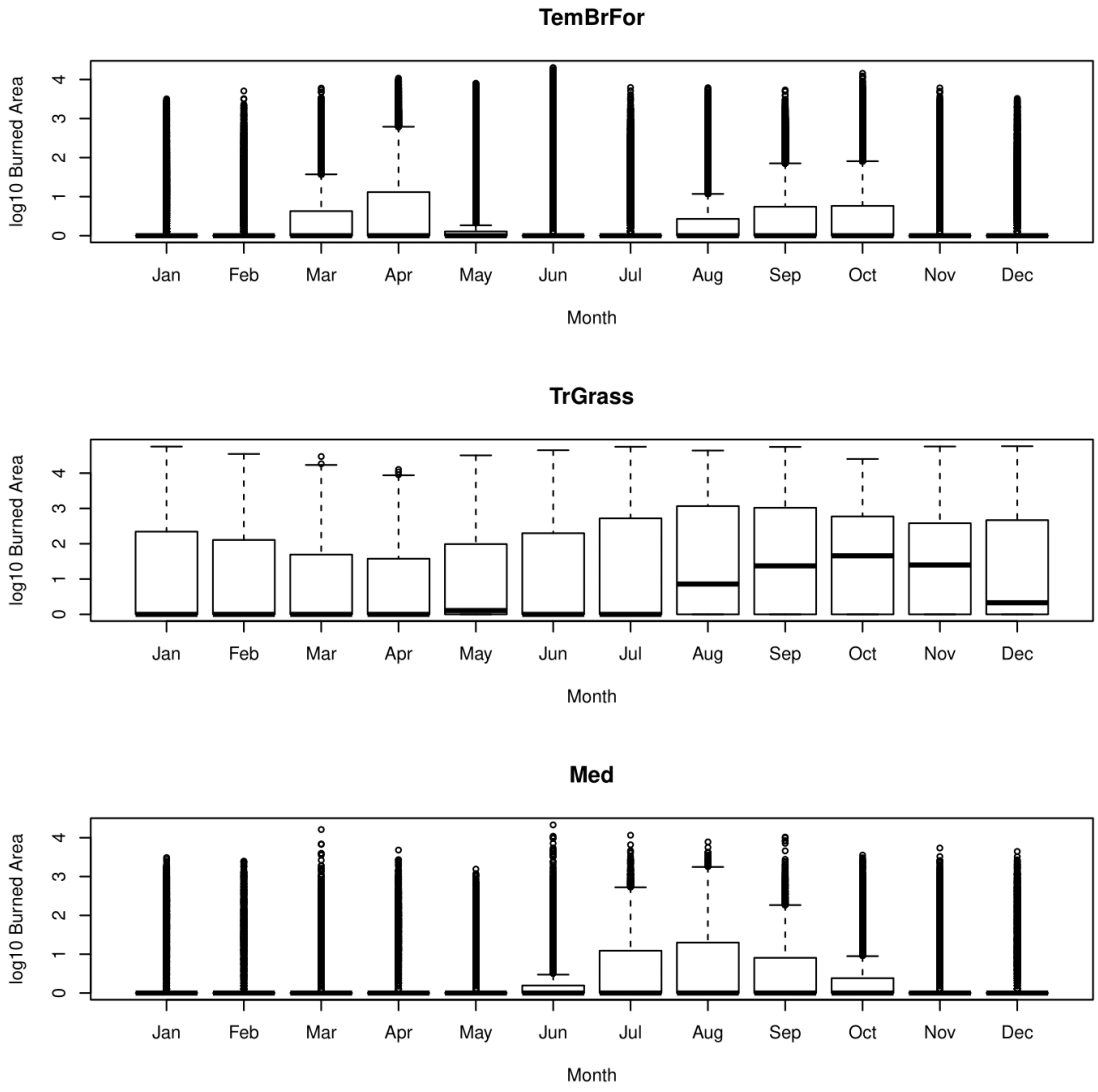

Figure 2.3: Monthly burned area (BA) distribution (ha, in $log10$ scale) in *TemBrFor*, *TrGrass* and *Med* biomes (Table 2.2), representative of three distinct types of fire regimes. *TemBrFor* exhibits a bimodal annual cycle with two marked BA peaks in spring and fall. *TrGrass* has an unimodal-type fire season (peaking in fall), but large fires can occur at any time throughout the year when the global biome is considered as a whole. Finally, the Mediterranean region (*Med*), characterized by a unimodal annual cycle with a strong peak in boreal summer and more rare fire events outside this season.

CHAPTER 3

Methods

3.1 Clustering Methods

Clustering techniques are unsupervised learning algorithms that group the data in different clusters trying to put together similar data. As a result, these algorithms assign one (and only one) group to each data point. The number of clusters to form is a non-trivial decision that the user must make taking into account some validation metrics. In this work, Gaussian Mixtures is the main clustering method used.

3.1.1 Gaussian Mixtures

Gaussian Mixtures are based on the fact that the data points belong to a probability distribution which is a weighted sum of K multivariate Gaussians (Murphy, 2012). This clustering technique is implemented in the *mclust* package (Scrucca et al., 2016). Gaussian Mixtures adopt the following form:

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k),$$

where x_i is one observation, θ represents the parameters of the models (π_k, μ_k, Σ_k), π_k is the mixing weight verifying $\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$ and μ_k and Σ_k are the mean vector and the covariance matrix of each of the Gaussians for $1 \leq k \leq K$. The model produces K clusters, each of them associated to one different Gaussian distribution —we have chosen this clustering model because fire seasons tend to have more or less the shape of a Gaussian or the sum of two of them.—

The model tries to find the value of the parameters $\theta = (\pi_k, \mu_k, \Sigma_k)$ that maximize the negative log likelihood of the observed data

$$l(\theta) = \log(p(x_i|\theta)).$$

In other words, it tries to obtain the parameters that make the observations more likely to belong to the distribution. As it could be very difficult to find the maximum of l, Gaussian Mixtures use the Expectation Maximization (EM) algorithm because it increases monotonically l. EM is an iterative method which has the following steps:

- **E step:** For each point x_i, calculate the probability of belonging to the cluster/distribution k using the expression

$$r_{ik} = \frac{\pi_k p(x_i|\theta_k^{(t-1)})}{\sum_j \pi_j p(x_i|\theta_j^{(t-1)})},$$

  where $\theta_k^{(t-1)} = (\pi_k, \mu_k, \Sigma_k)$ for $1 \leq k \leq K$ in the $t-1$ realization of the EM algorithm and $1 \leq j \leq K$.

- **M step:** Recalculate the value of the parameters, where N is the number of observational points:

  1. $\pi_k = \frac{1}{N} \sum_i r_{ik}$
  2. $\mu_k = \frac{1}{r_k} \sum_i r_{ik} x_i$
  3. $\Sigma_k = \frac{1}{r_k} \sum_i r_{ik}(x_i - \mu_k)(x_i - \mu_k)^T$.

3.2 Predictive Techniques

We have considered for this work three type of data mining techniques to develop predictive models of BA. First, we have used linear models as a benchmark because of their simplicity (and their reasonably good performance regression-like problems). Besides, we have also considered random forests and k-nearest neighbours.

3.2.1 Linear Models

An advisable way to start dealing with a problem like ours is to begin with the easiest method and, if needed, move to more complicated ones. For this reason, we have first considered as benchmark a simple linear regression.

This technique just tries to find the straight line that best fits our data. Its objective is to find the coefficients α_i which produce the minimum value of the mean square error (MSE) between observed and predicted data. Hopefully, the problem has an analytic solution that is given by the expression

$$\alpha = (X^T X)^{-1} X^T y,$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, $y = (y^{(1)}, y^{(2)}, \ldots, y^{(n)})$ is the observed data and

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \ldots & x_m^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \ldots & x_m^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \ldots & x_m^{(n)} \end{pmatrix}$$

is the characteristics matrix whose rows are observations and whose columns are the variables that we will use for the prediction.

3.2.2 Random Forest

Bagging is a learning paradigm which is based on the idea that combining many weak models can produce a strong one which yields improved stability and predictive power. Random forests are a particular type of bagging technique which combine many individual regression trees. A regression tree is a simple model that is used to predict a continuous target variable. It is made up of nodes, branches and leaves where each node represents a test on an attribute, each branch corresponds to an attribute value and each leaf (terminal node) represents a final class. Tree building algorithms evaluate an attribute according to its power of separation which is given by the Residual Sum of Squares (RSS) (Eq. 3.1):

$$\sum_{j=1}^{J} \sum_{i:\, x_i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{3.1}$$

where $R_1, R_2, \ldots, R_J$ are the regions in which the characteristics space is divided, y_i are the points of each region and $\hat{y}_{R_j}$ is the average of all the points of the region. In other words, the new division in the tree, which will result in two new nodes, is the one that reduces more the value of the RSS. However, trying all the possible combinations is computationally expensive, so we use recursive binary separation.

In a first step, we consider the whole region of the characteristics space and we try to divide it in two parts finding the predictor j and the separation threshold s that minimize Eq. 3.1 for $J = 2$:

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

where $R_1(j, s)$ and $R_2(j, s)$ are the following sets:

$$R_1(j, s) = \{X \mid X_j \leq s\},$$
$$R_2(j, s) = \{X \mid X_j > s\}.$$

Then, we apply the same idea to the new regions R_1 and R_2 and so on until one stopping criteria is achieved (typically one of the following: error reduction too small versus tree complexity, maximum depth achieved or too small sub-samples in some region).

Trees often suffer from overfitting. In order to avoid it and obtain a better and robust model, random forests combine many different regression trees. The idea is to select m subsamples of the training data of the same size using bootstrapping and to grow one regression tree for each subsample. Hence, we will end up with m different trees and the final prediction will be a combination of the m predictions given by each single tree, typically the average of all of them for regression problems. However, if we use the same predictors for all the trees, we will get very similar models, reducing the generalization ability of the random forest. Consequently, for each regression tree we will only use *mtry* number of predictors that are randomly selected.

In this project, we have used the random forest implementation which is available in the *caret* and *randomForest* packages. Based on a proper cross-validation scheme (Sec. 3.3.1) we have optimized the number of predictors (*mtry*), the number of trees (*ntree*) and the maximum number of terminal nodes (*maxnodes*) for our random forests based on the correlations attained.

3.2.3 k-NN

k-Nearest Neighbours (k-NN) is a supervised learning technique. Given the value of k and a point x_i, the algorithm chooses the k points $x_1, x_2, \ldots, x_k$ that are more similar (in the sense of the some distance, for example the Euclidean one) to x_i. The final prediction for x_i will be a combination (typically the average) of $y_1, y_2, \ldots, y_k$, i.e., the value of the target variable in each point $x_1, x_2, \ldots, x_k$.

3.3 Validation Framework

In order to measure the capability of generalization of our predictive techniques and to reduce the probability of overfitting, a suitable validation framework must be defined.

3.3.1 Cross-validation

First, as we want to develop models with a good ability to generalise, cross-validation should be considered. Otherwise, our models may suffer from overfitting, i.e., they might be very accurate with train data but they will be unable to make correctly extrapolate to unseen test data.

The easiest approach for cross-validation is *hold out*, which splits the full data available up in two separated groups. The first one, which includes the majority of the data points (typically 75% or 80% of the whole data), is used for training the model whereas the other part is used to test the model quality. However, deciding how the data separation must be performed is a non-trivial issue because different partitions might lead to different results. In this project we have used a random partition containing the 70% of the total dataset for training, which allows to optimize the parameters of both the random forests (*mtry*, *ntree* and *maxnodes*) and the k-NN technique (k).

More sophisticated approaches for cross-validation consist of dividing the whole dataset in more than two groups. This is called k-fold cross-validation, where k is the number of partitions. For each $1 \leq i \leq k$ we build a model using the j-th group (fold), $j \in \{1, \ldots, k\} \setminus \{i\}$, and then, we make a prediction for the i-th group.

In this work, we have used *leave-one-out* cross-validation, which is a k-fold with $k = n$, where n is the number of points we have. That is to say, for each $1 \leq i \leq n$ we train with all the data except the i-th point and we make a prediction for this point. This method is only used when there are few points as it is our case. Otherwise, it is computationally expensive. We used it to train all the predictive models.

3.3.2 Validation Metrics

To assess the quality of our predictive models in a comprehensive way, we have considered several metrics which allow to evaluate different forecast aspects. These metrics are described next, using the o (p) letter to represent the observations (predictions).

- **Bias:** It represents the difference between the average of predicted and observed values, so the best value we could get is 0 (unbiased model). For direct comparison across different clusters, we express it as a percentage —with respect to the standard deviation of the observations, σ_o,— where $\overline{o_i}$ ($\overline{p_i}$) is the average of the observations (predictions):

$$bias = 100 \left(\frac{\overline{p_i} - \overline{o_i}}{\sigma_o} \right).$$

- **Ratio of variances:** It provides a simple representation of the fraction of observed variance explained by the predictions, being 1 the best possible achieved value:

$$RV = \frac{var(p)}{var(o)}.$$

- **Correlation:** It measures how well the predicted time-series follow the corresponding observations in a range between -1 and 1, where the latter is the desired score. We use the Pearson's coefficient, being n the number of observations:

$$r = \frac{\sum_{i=1}^{n}(o_i - \overline{o_i})(p_i - \overline{p_i})}{\sqrt{\sum_{i=1}^{n}(o_i - \overline{o_i})^2}\sqrt{\sum_{i=1}^{n}(p_i - \overline{p_i})^2}}$$

- **Lower/middle/upper tertile accuracy:** It gives a measure of the number the times the lower/middle/upper observed tertile is correctly forecast, being 1 the best possible score:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Number of predictions}}$$

- **Tertile accuracy:** It gives a measure of the number of times the observed tertile is correctly forecast.

  Note that, in the last two metrics the tertiles are independently computed for the actual and the predicted time-series, which makes them bias-insensitive.

CHAPTER 4

Results and Discussion

4.1 Clustering

Clustering was a crucial step in this project. As Gaussian Mixtures may be a non-deterministic model when we are working with many points, several attempts were made for each biome. The final choice (Fig. 4.1) was done according to the best obtained Bayesian Information Criteria (BIC), as suggested in Fraley (1998), after several preliminary exploratory data analyses. For instance, the influence of outliers was first analysed. Large BA pixels unduly influence the clustering results and tend to accumulate most of the clusters despite the shortage of these pixels. As a result, BA data was *log*-transformed prior to clustering in order to reduce the large BA magnitude differences between close pixels, leading to a more robust grouping of pixels attending to their similarity in annual cycle shapes, and to a more homogeneous spatial distribution of types (see Fig. 2.1).

Despite *log*-transformation, a non-optimal classification was found due to the large differences of total BA across regions of the world —pixels with low values of BA were grouped together in spite of their differing annual cycles.— In order to obtain more robust results, we undertook a stratified clustering based on a global biome classification (Olson et al., 2001, see Sec. 2.2). This provides a suitable basis for separately analysing the data attending to similar biophysical features (fuel types, climate ...), with the double advantage of providing a better discrimination of fire season types within each region and dramatically improving the computing times of the Gaussian Mixture algorithm, due to the reduction of within-biome BA

variability on the one hand (see e.g. Fig. 2.3), and total sample size on the other. In order to avoid too many clusters, an additional stopping criterion to BIC was imposed, consisting in setting a minimum cluster size of 5% of the available points within the biome.

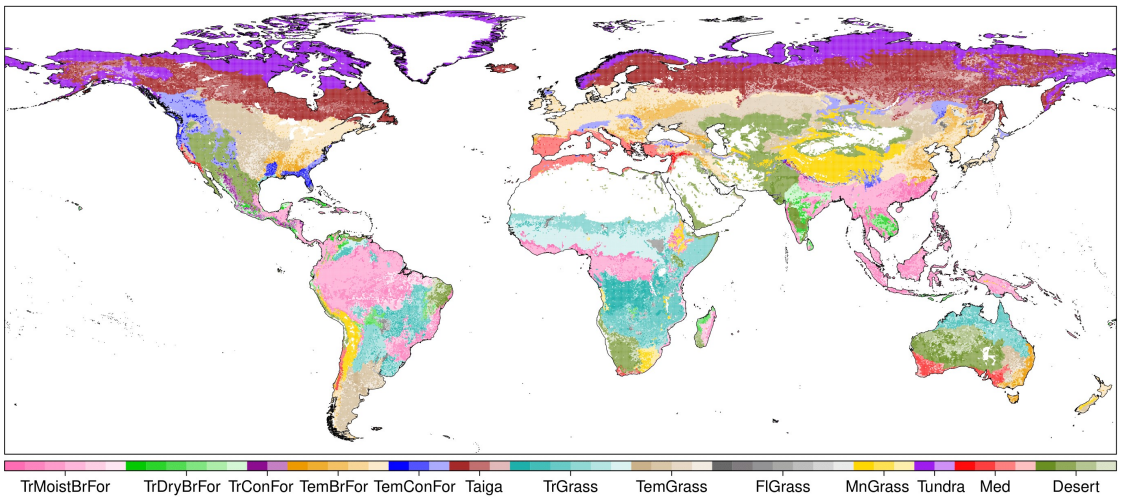

Figure 4.1: Fire season types resulting after the biome-based clustering. Each color corresponds to a different biome. Within each biome, the different clusters are indicated by different color saturation levels.

Finally, we obtained between 2 and 7 clusters per biome, making a total of 55 clusters which exhibit a spatially coherent distribution across the globe (Fig. 4.1). The requirement of having a minimum of 5% of the pixels contained in the biome to form a cluster avoided the appearance of too many, meaningless clusters per biome —the BIC criterion alone was not enough in this respect.— This ensured a good compromise between diversity and an adequate grouping of similar fire season types (i.e., same start/end and timing of the BA peak). As an example, using only the BIC criterion the Tundra biome yielded around 40 different fire season types of very low frequency, while the 5% criterion differentiates just two distinct, spatially coherent types that reflect a latitudinal gradient within the Arctic region (see Fig. 4.1). Further details for each biome are provided in the dedicated notebook of the supplementary material[1] (Sec. 5.3).

[1] `https://github.com/MarcosVM98/TFM/blob/master/notebooks/Definitive_Clustering_v2.ipynb`

4.1.1 Fire Season

The fire season is typically defined by the months of the year encompassing the bulk of BA. In particular, in this work we have defined the fire season as the months of the year summing, in average, more than 80% of the total annual BA in each cluster, following the European Forest Fire Information System (EFFIS, `https://effis.jrc.ec.europa.eu/`) criterion (Jesús San Miguel Ayanz, *pers. comm.*). Note that the average period used here is determined by the fire data availability at the moment of the analysis (Jan 2001-Apr 2020, Sec. 2.1). When the months encompassing this period are not consecutive (i.e., two differentiated BA peaks exist throughout the year) a "main" fire season and a "secondary" fire season are distinguished, yielding a bimodal fire seasonal cycle that is often signaled as a human-induced feature (see e.g. Benali et al., 2017) Therefore, in order to characterize the annual cycle for each pixel, for each month of the year we pick the 75^{th} percentile of BA along the full time series, yielding one BA value for each month, from January to December, that we refer to as the "annual cycle" of BA (see e.g. Fig. 4.2). Then, we pick the first j months that sum more than the 80% of the total burned area of the vector and we take the period(s) of consecutive months. The analysis of this annual cycle for each pixel allows for the differentiation between unimodal and bimodal cycles, that are afterwards treated separately in subsequent analyses (see Fig. 4.3).

For better adaptability of our approach, both parameters, BA fraction and reference percentile (80% and 75^{th} respectively in this study) can be arbitrarily modified by the user, as indicated in the reproducibility notebooks (Sec. 5.3). Note that we tested several different percentile thresholds before opting by the 75^{th} (notably the median and also higher percentiles). The final choice was partly guided by our own "expert" knowledge, yielding very consistent results for well known regions like the Mediterranean (Fig. 4.2), where wildfires are a key natural hazard. In this region, the median produced inaccurate fire season results. Likewise, the mean proved to be too sensitive to outlying observations (i.e. exceptional BA records attained in some years).

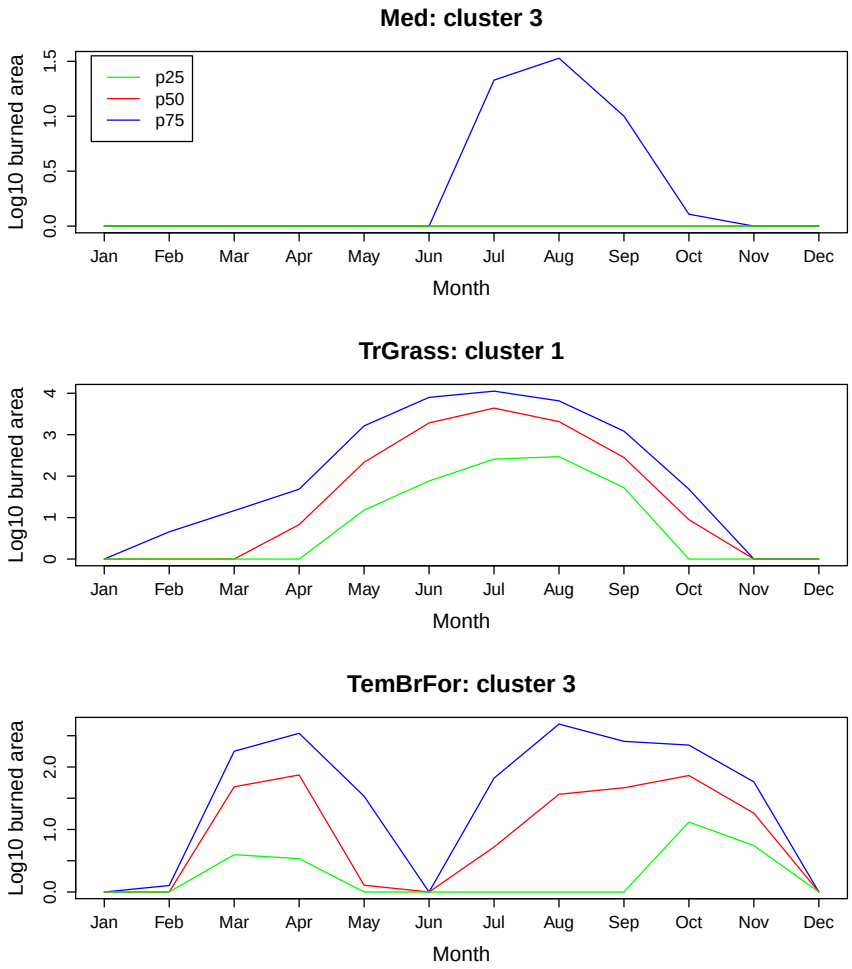

Figure 4.2: Burned area distribution in 3 illustrative clusters. Green, red and blue lines represent the 25^{th} percentile, the median and the 75^{th} percentile, respectively. The blue ones (75^{th} percentile) are used to obtain the fire seasons.

Our analysis across the globe reveals that sizeable areas of Central and Northern Europe, East Coast of North America and the Amazon Basin exhibit very low BA records (the sum of the *log* BA of all the pixels within the clusters during the fire season is less than 0.2) and that in the East of Europe and Asia there are regions with *bimodal* fire seasons —two separated periods of fire activity during the year,— as shown in the third panel of Fig. 4.2. Finally, *unimodal* fire seasons —one single main period of fire activity during the year— like the first and the second of Fig. 4.2 are distributed all over the world: the Sahel, prairies of central North America, East and South of South America or the South of Europe (EU-Med region). Overall, our results are consistent with previous studies analysing the fire season globally (Benali

et al., 2017) relying on a different statistical parametric approach operating pointwise. This finding highlights the advantage of the automated clustering method presented here, that can be easily applied to multiple spatial and temporal scales with minor parameter modifications. Furthermore, the approach can be successfully used to automatically differentiate unimodal and multimodal fire season cycles, as revealed by the comparison of our results with other regional studies like Ameijeiras-Alonso et al. (2019) over eastern Europe dry grasslands.

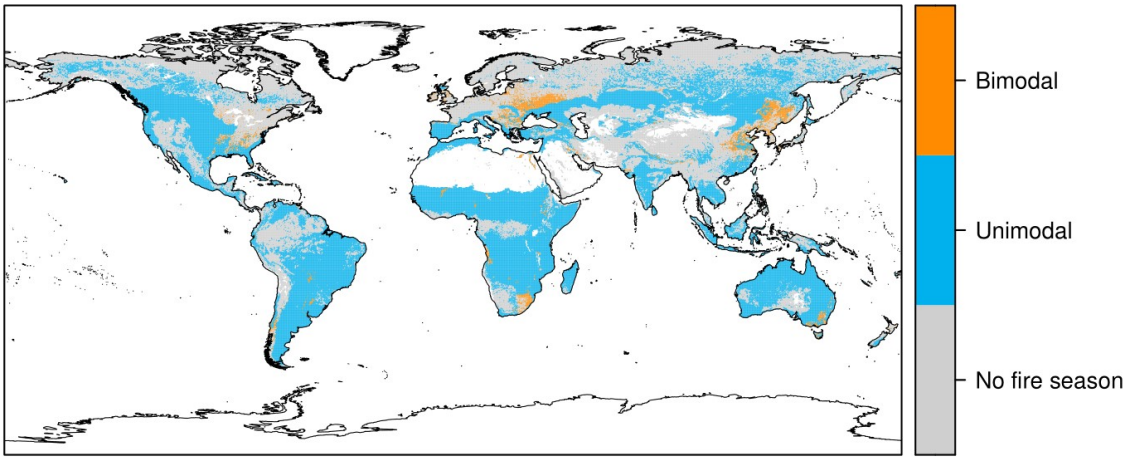

Figure 4.3: Type of fire season (unimodal or bimodal) as automatically calculated from the BA annual cycles at the pixel scale. Blanked areas correspond to masked pixels following the 10% fraction of burnable area criterion (Sec. 2.1).

Therefore, our results depict an unambiguous pattern of fire seasonality consistent with previous studies, able to characterize the timing of BA peaks as well as the duration and shape of the pixel-scale fire season relying on a fully automated procedure that can be tuned by the user through the manipulation of a few simple parameters in order to accommodate fire datasets of varying nature, spatial extent and spatial and temporal resolution.

4.2 Predictive Models

Taking into account the biome classification and the fire season calculation described in Sec. 4.1.1, we developed empirical predictive models using as the only predictors the climate teleconnection indices introduced in Sec. 2.3.

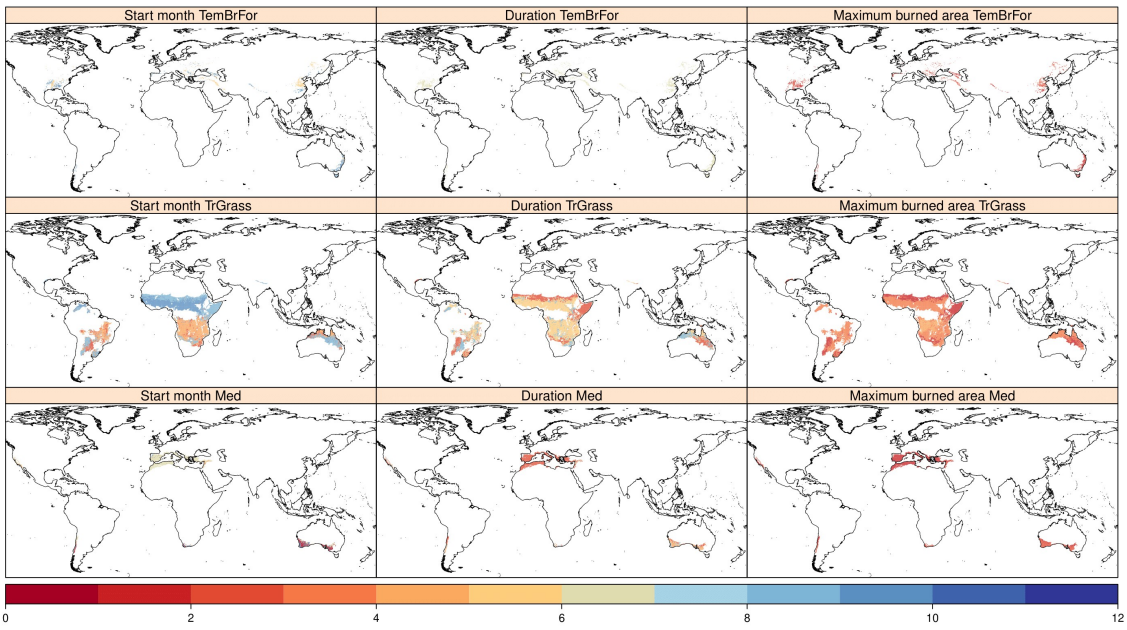

Figure 4.4: Start month, duration of the fire season and maximum burned area of the peak month (in $log10$ scale) of the unimodal fire seasons of the *TemBrFor*, *TrGrass* and *Med* biomes.

4.2.1 Correlation Analysis

Before building our predictive models we performed an exploratory correlation analysis to assess the existing degree of association between different climate indices. This was done to avoid working with redundant predictor variables, which might have a deleterious effect on our models, especially on linear regression ones. Alternative approaches such as the use of principal components would serve to the same purpose (see, e.g., Rodrigues et al., 2021), but at the cost of hindering model interpretability, so it was discarded.

A correlation matrix for all climate indices is shown in Fig. 4.6. It can be seen that ONI, SOI and El Niño 3.4 indices are highly correlated. Consequently, ONI and SOI were discarded hereafter. This choice was done based on a potential extension of the work presented here in which seasonal forecasting models could be used to implement an operational early warning system for fire activity. In this regard, the state-of-the-art seasonal climate models provide very accurate predictions of El Niño 3.4 index with a few months of anticipation (see, e.g. Manzanas et al., 2014).

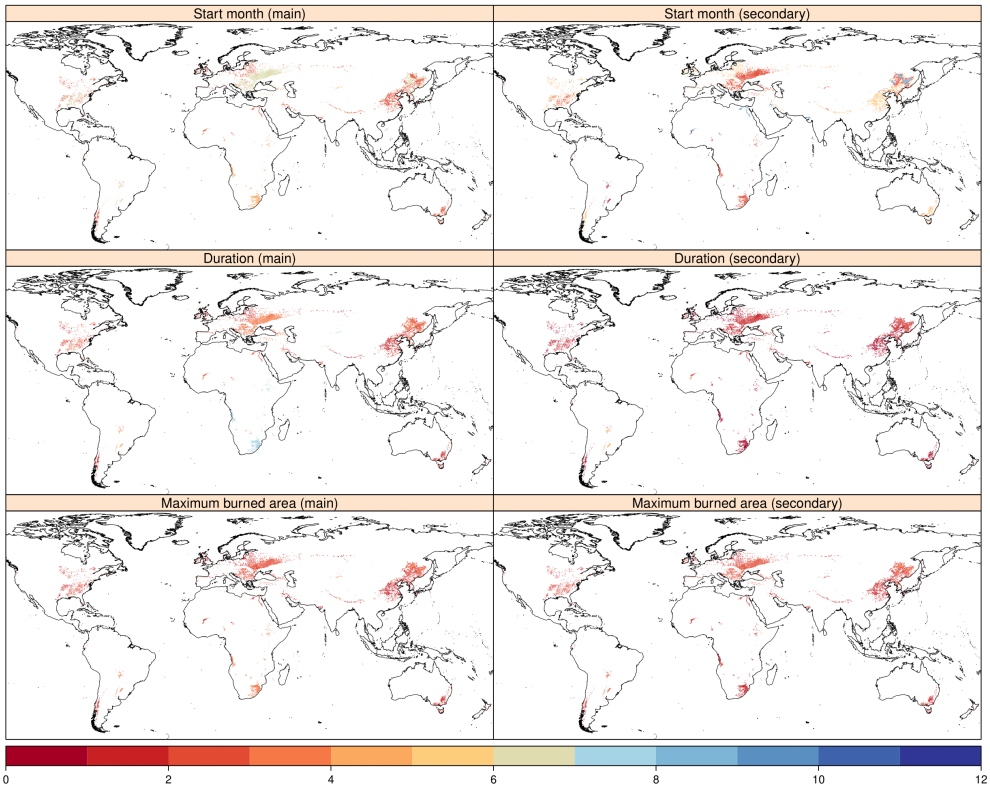

Figure 4.5: Start month, duration and maximum burned area (in $log10$ scale) of the global bimodal fire seasons, considering the "main" fire season (left) and the "secondary" fire season (right). See Sec. 4.1.1 for the definition of main and secondary fire seasons.

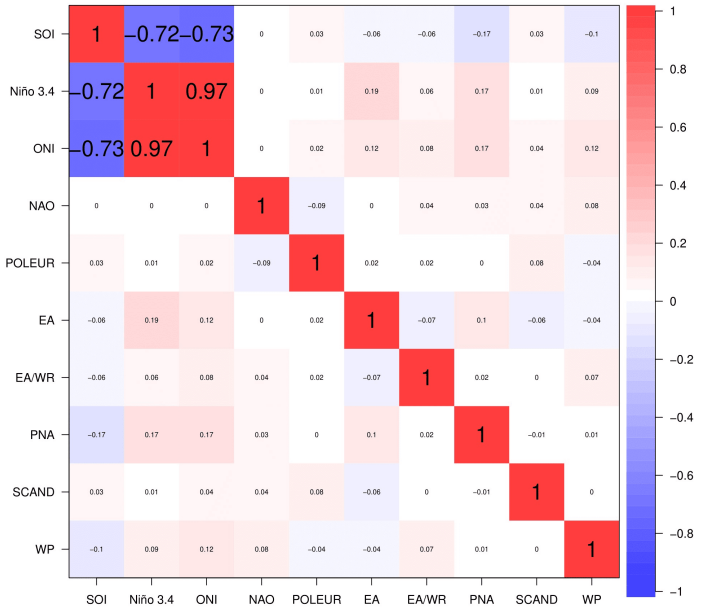

Figure 4.6: Pearson correlation coefficient between all climate indices considered. Red (blue) colors shows positive (negative) values.

Considering the final subset of uncorrelated teleconnection indices, we then computed the correlations of each index mean with total BA during the fire season, considering in this case only the unimodal fire season pixels.

The influence of the climate indices might not be immediate and their effects on BA might not be seen until some time had passed (see e.g. Rodrigues et al., 2021). Therefore, we have considered in our study time-lagged climate indices by 1, 2 and 3 months. For instance, if the fire season includes June, July and August, lag 1 refers to the climate index values of May, June and July.

Furthermore, as the possibility of autocorrelation due to trends in the data is a potential problem when comparing two time series, we use a common approach in crop research (Lobell et al., 2007) based on calculating the first-order time differences of predictor and predictand (we will refer this to as *deltas* hereafter); this is, we consider the year-to-year difference of the variable instead of the original variable values. As a result, in the delta approach the correlation is calculated between the year-to-year differences (between two consecutive fire seasons) of the total sum of the *log* of the BA and the averaged climate index. Furthermore, this approach is also used in order to enhance the climatic signal and to separate this from other confounding factors affecting fire (see e.g. Turco et al., 2014; Urbieta et al., 2015; Bedia et al., 2015, in the context of fire research).

To assess the effect of building or not on the delta approach, the correlation analysis was undertaken both for the year-to-year differences in BA and for the original BA values. The difference in the results using both approaches is illustrated in Figs. 4.7 and 4.8, which show the correlation per cluster between total BA (in *log* scale) and two illustrative climate indices (PNA and SCAND) within the fire season for different time lags. These results exhibit a reinforced signal when the delta approach is used, highlighting its adequacy to model the fire-climate relationship.

Our results show that the PNA index obtains stronger correlations in the centre of South America, Indonesia and the south of Africa for lags 1, 2 and 3 than for lag 0. In addition, the use of deltas tends to strengthen the correlation, as seen in southern Africa for the PNA or around the Mediterranean basin for the SCAND. In general, the PNA index exhibits correlations over 0.6 in parts of South America, Africa and Indonesia and SCAND presents significant negative correlation in the Iberian Peninsula, where Rodrigues et al. (2021) also obtained similar results but using the Fire Weather Index as predictand. The remaining climate indices also obtain significant correlations in different parts of the world. Further details for

each index are provided in the dedicated notebook of the supplementary material[2] (Sec. 5.3).

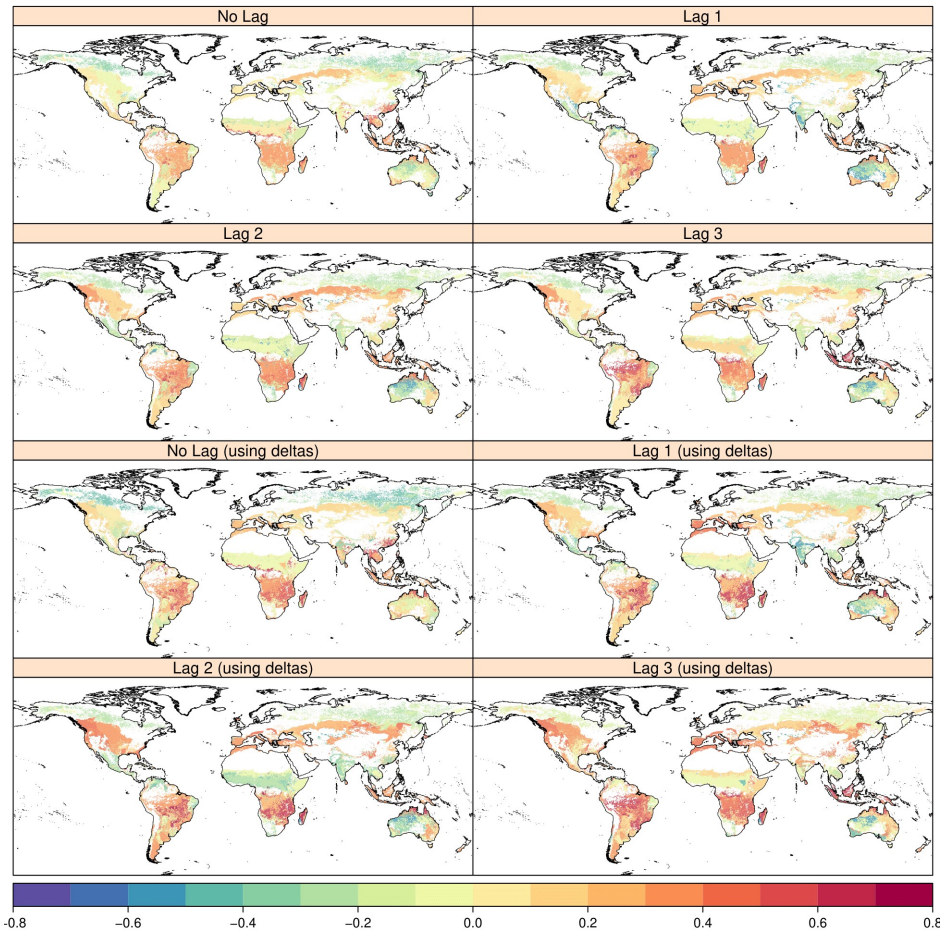

Figure 4.7: Panels 1 to 4 (starting from the left-upper corner, and from left to right): Pearson correlation between the mean PNA index and the sum of the *log* of BA in the fire season, for different time lags. Panels 5 to 8: As panels 1 to 4, but using the delta approach. In 95% of the cases, correlation coefficients outside the $[-0.46, 0.46]$ range are statistically significant at a 95% confidence level. White areas are either masked or do not have a fire season.

4.2.2 Predictive Models

With the idea of finding the best predictive model for BA, we have considered three different data mining techniques in this work: linear models (only with certain pre-

[2] `https://github.com/MarcosVM98/TFM/blob/master/notebooks/Correlation_Per75_with_deltas_v2.ipynb`

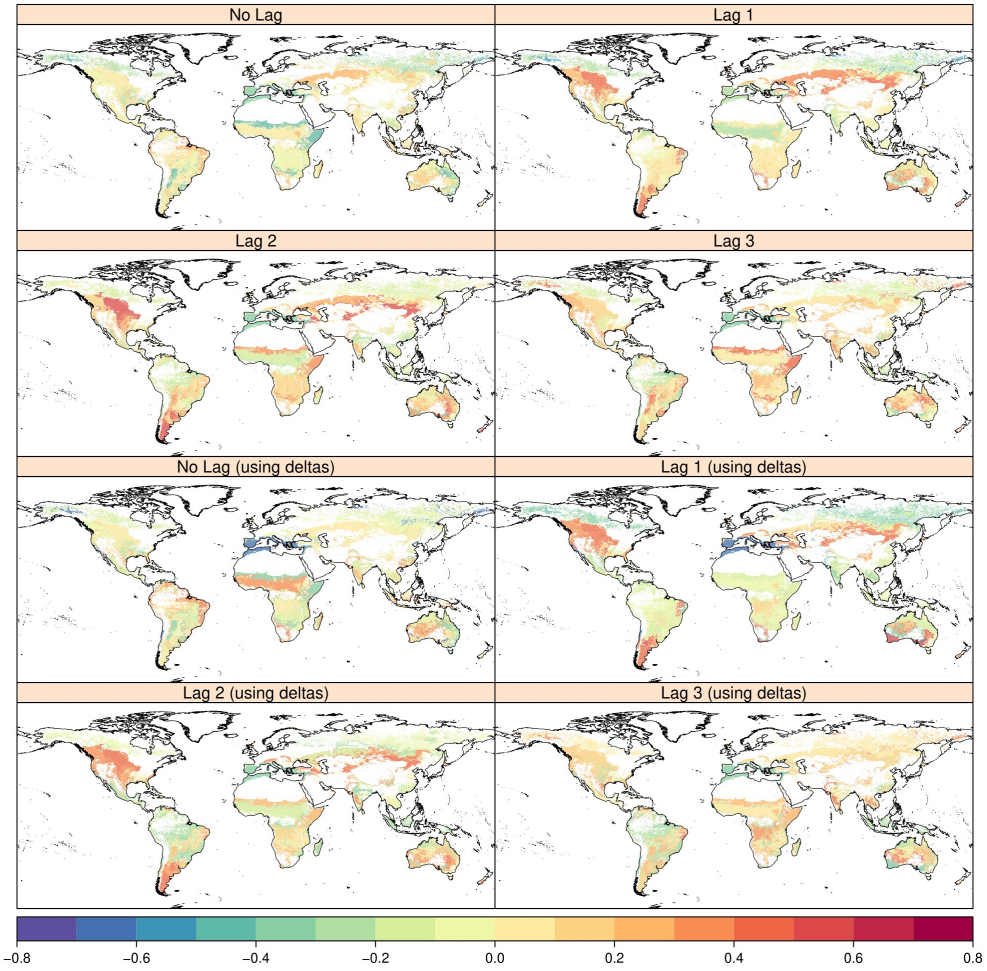

Figure 4.8: As Fig. 4.7 but for the SCAND index.

dictors), random forests and k-NN (the last two use all the available predictors). Furthermore, for each technique we have built six different models considering different time lags of the climate indices used as predictors. We have one model per each lag between 0 and 3, then another that considers predictors with lags 1 to 3 (combined) and finally one more which uses all possible lags (i.e., 0, 1, 2 and 3). The target variable is the delta between the total of the *log* of the burned area in the cluster during two consecutive fire seasons as described in Sec. 4.2.1, and the predictors are the deltas of the averaged climate indices.

The exploratory correlation analysis described in Sec. 4.2.1 served to the purpose of finding adequate predictors for the linear models. That is, only those climate indices exhibiting a significant correlation with BA were used as predictors in this type of models. However, all the available climate indices (excluding ONI and SOI)

were considered as potential predictors for the case of random forest and k-NN models. As a result, there are several clusters which do not have a linear model for certain lags and there are also linear models which use one single predictor, as shown in Figs. 4.9 and 4.10 (the missing linear models are indicated by the 0 number of predictors category).

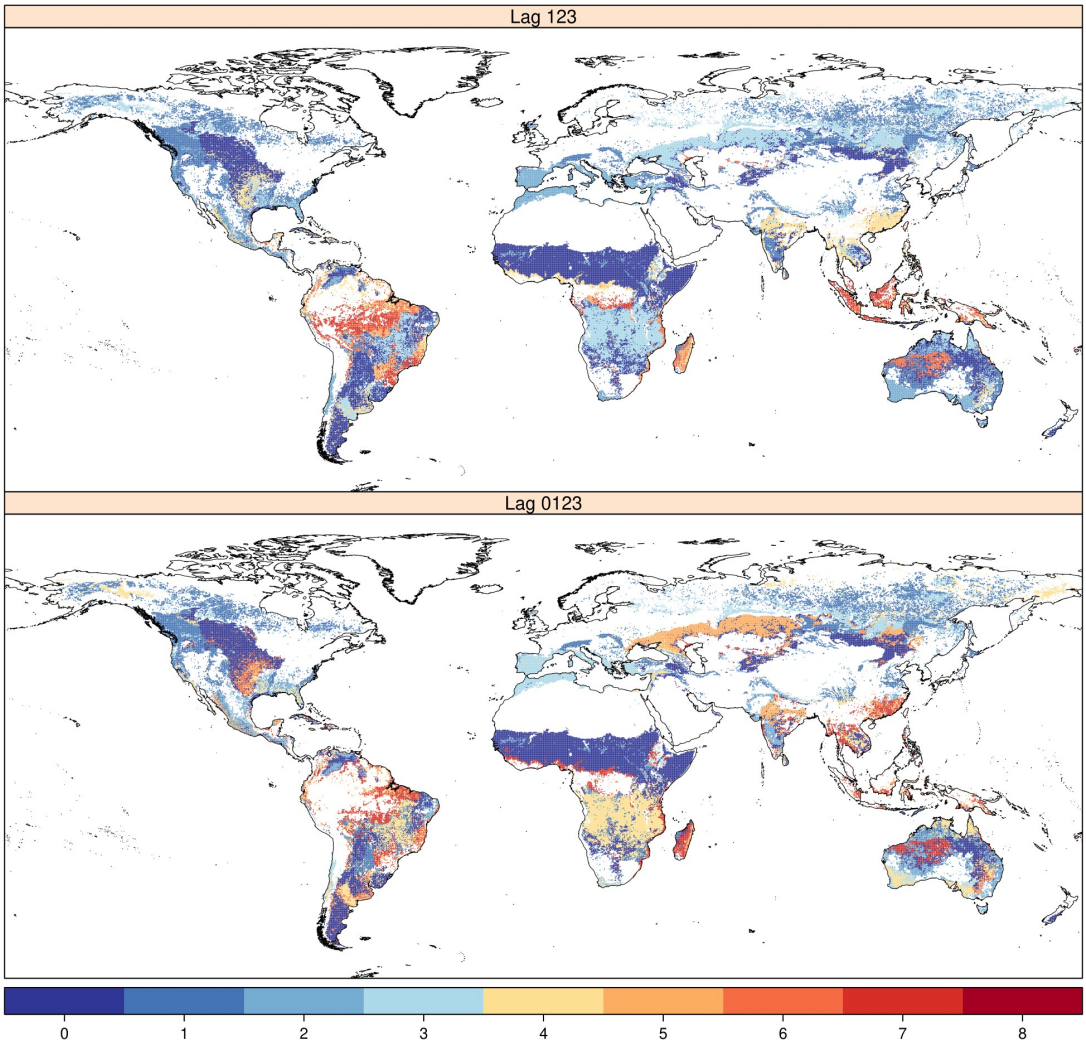

Figure 4.9: Number of predictors considered in the linear models using predictor information which combine different time-lags, at the cluster-level.

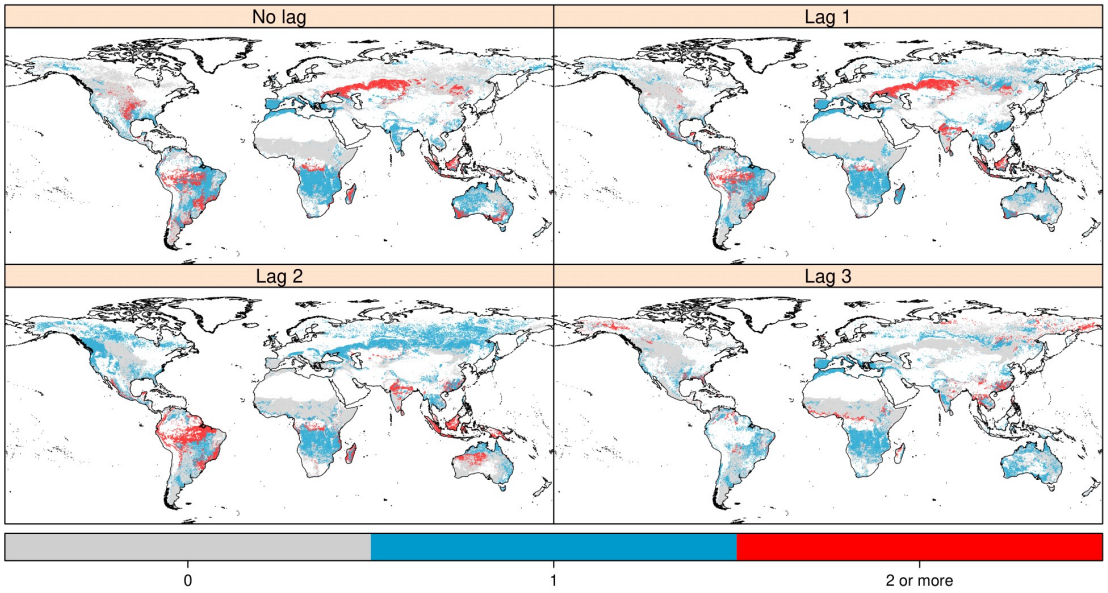

Figure 4.10: Number of predictors considered in the linear models using predictor information at specific time-lags, at the cluster-level.

The main parameters for the random forests and the k-NN techniques (number of trees and number of neighbours, respectively) were optimized for each cluster and lag according to the value of the correlation attained under the cross-validation scheme described in Sec. 3.3.1. Figs. 4.11 and 4.12 represent the optimum value obtained for the *ntree* and k parameters (number of trees and number of neighbours, respectively) for each cluster and time lag.

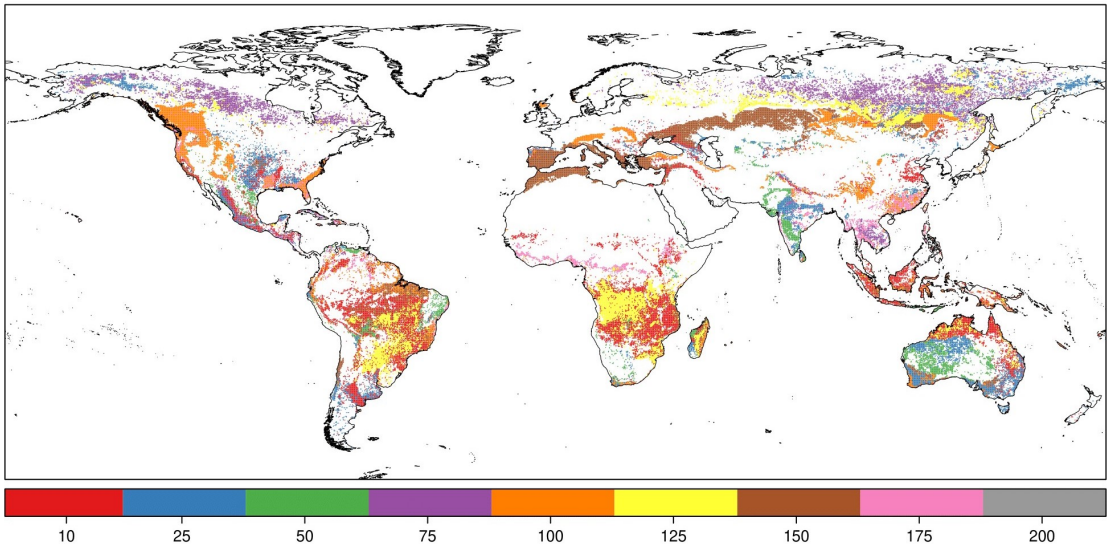

Figure 4.11: Optimum number of trees for the random forest technique at time lags 0 to 3 (*rf_0123*) for each cluster, according to the correlation attained in cross-validation mode. Clusters that do not have a linear model are not represented.

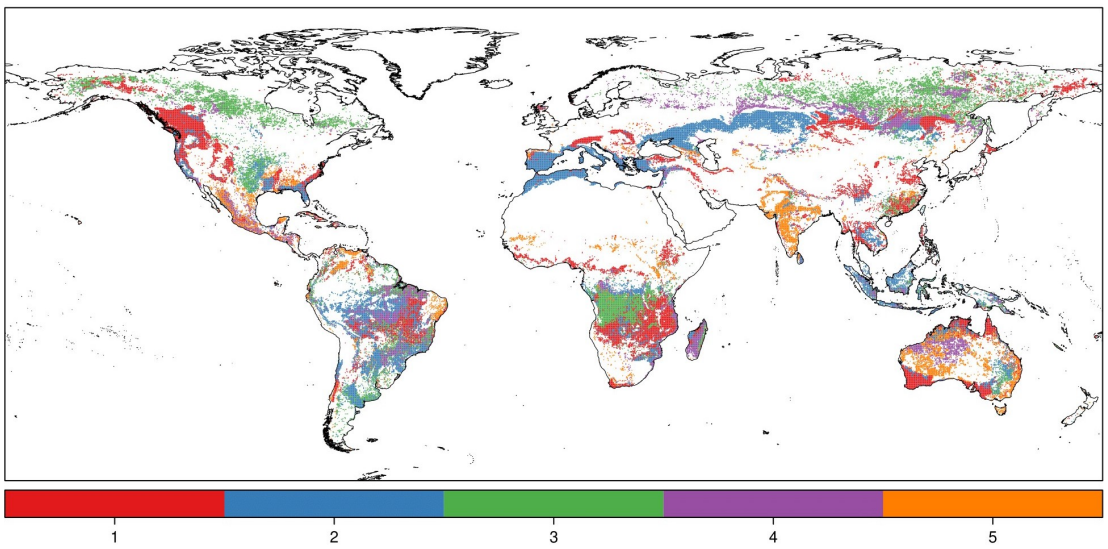

Figure 4.12: As Fig. 4.11 but for the number of neighbours parameter of the *kNN_0123* models.

According to Fig. 4.13, which shows violin plots[3] that allow to summarize our results in terms of the validation metrics explained in Sec. 3.3.2, providing a good representation of the performance of all the models applied for each specific technique and lag, linear models are the most competitive ones in most of the cases, specially for the third tertile accuracy and for correlation, which are key with regards to a potential use of this methodology in an operational forecast context. This has major importance because it suggests that our climate indices and the BA are linked through simple, mostly linear relationships. In contrast, k-NN models tend to be biased and have low correlation while random forests are too conservative since they have small ratio of variance and also they obtain poor results in the upper tertile accuracy. In addition, even the worst linear models obtain better results in almost every validation metric, except for the middle tertile accuracy, that is usually the more difficult to predict for the linear models. As expected, linear models with more predictors perform better; in particular, the best results are found when predictors of lags 0 to 3 are included. However, this pattern, which also appears for the k-NN models, is not clearly shown in random forests because the ones having more predictors (rf_123 and rf_0123) get worse results in bias and ratio of variance, though they improve their tertile accuracy and correlation. We ought to note that half of the *lm_0123* models get correlations and third tertile accuracy over 0.5. Fig. 4.14 shows a comparison between the observed time-series and predicted ones — obtained with the *lm_0123* linear model— for four illustrative clusters (note that the top row correspond to the unimodal clusters shown in Fig. 4.2). Noticeably, correlations above 0.5 are attained in all cases, reaching values of about 0.8 in the clusters shown in the bottom row (which have been selected with illustrative purposes).

[3]Violin plots are similar to box-plots but they also include an estimation of the probability density of the data.

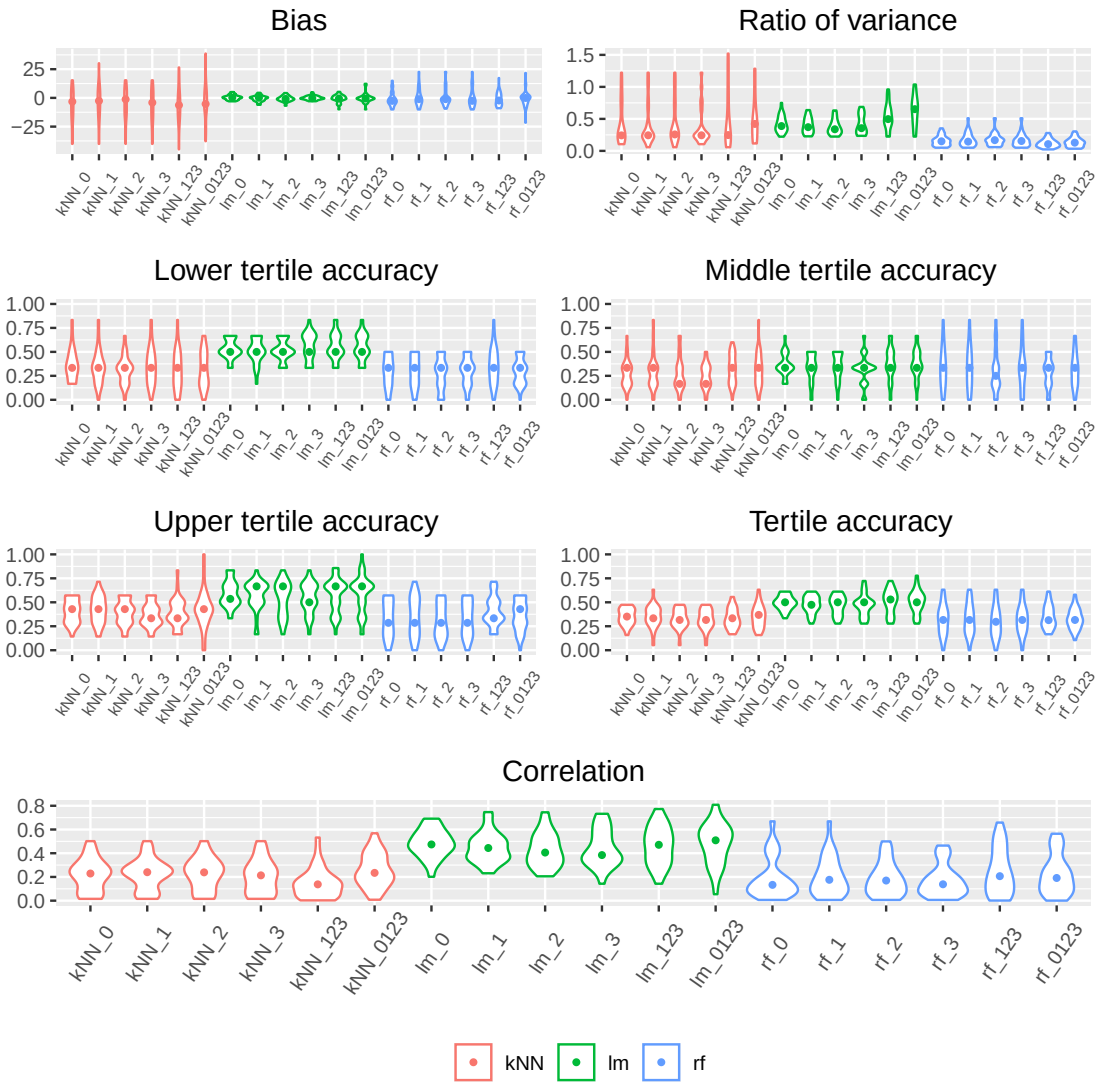

Figure 4.13: Summary of the results obtained (each color correspond to a different predictive technique) in terms of different validation metrics. Correlation is in absolute value and bias is expressed as a percentage with respect to the standard deviation of the observed values. Models (random forests and k-NN) in clusters and lags where there are not significant correlated indices are not considered. The point of each violin represents the median across all considered clusters.

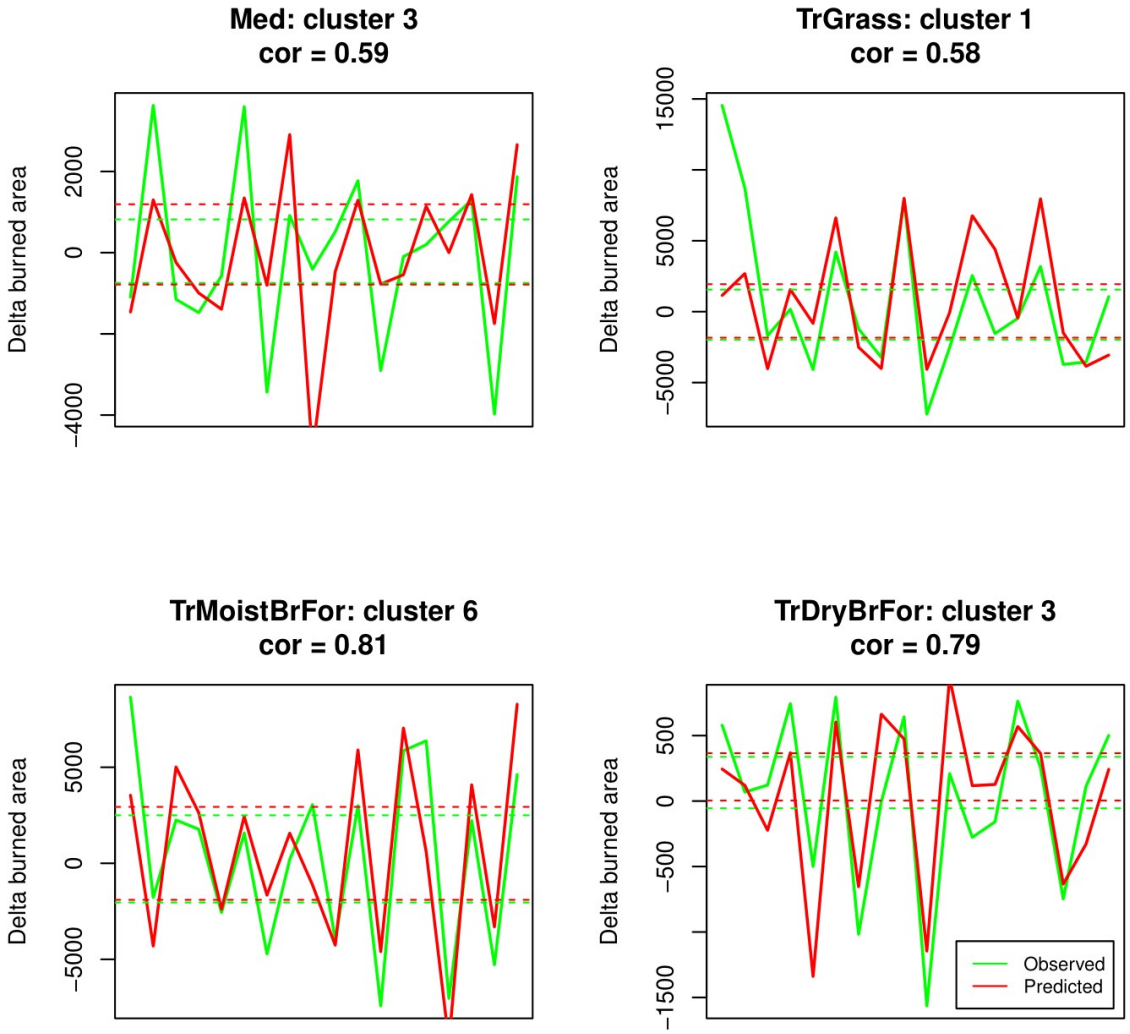

Figure 4.14: Predicted and observed time-series for delta differences of the BA (in *log* scale) in four illustrative clusters. Numbers above each panel show the Pearson correlation between observations and predictions. Horizontal lines in green (red) represent the tertiles of the observations (predictions).

CHAPTER 5

Main Conclusions and Future Work

5.1 Main Conclusions

Next, we briefly expose the key conclusions drawn from this work:

- We have introduced a new methodology based on Gaussian Mixtures which has proved successful in grouping global land pixels according to their fire season, properly distinguishing between unimodal and bimodal fire annual cycles. Overall, our results agree with previous studies using parametric tests for the same purpose. Our results depict an unambiguous pattern of fire seasonality consistent with previous studies, that is able to characterize the timing of BA peaks as well as the duration and shape of the pixel-scale fire season. The main advantage of the method presented is that it is a fully automated procedure that can be tuned by the user through the manipulation of a few simple parameters in order to accommodate fire datasets of varying nature, spatial extent and spatial and temporal resolutions.

- We have demonstrated that working with year-to-year differences (referred to as deltas) instead of actual values allows to improve the predictability of the burned area, strengthening the empirical links between this variable and the global climate teleconnection indices used as predictors. Moreover, the use of lagged climate indices unveils the existence of retarded links with fire activity in various parts of the globe, which helps to increase the performance of our predictive models, suggesting a promising potential of this approach for

its application within an operational fire prediction context at seasonal time scales.

- Even though we have considered more sophisticated data mining techniques (random forests and k-NN) our results evidence that parsimonious linear models, specially those which consider current and lagged predictors (lags 0 to 3), are the most appropriate for this task. This suggests that the relationships that link global climate teleconnection indices and burned area in many parts of the globe are simple and mostly linear.

- Despite our predictive models rely solely on climate teleconnection indices as predictor information and the sample size available for training was limited, we have reached high levels of skill to predict the BA for the upcoming fire season in various parts of the world, in particular for the majority of clusters included in sensitive world areas –either by their conservation value and/or for being densely populated regions– such as the Tropical and Subtropical Moist Broadleaf Forests, Tropical and Subtropical Dry Broadleaf Forests or the Mediterranean biomes, among others.

5.2 Future Work

We outline next some future activities which constitute the natural continuation of the current work:

- Apply the methodology proposed here to other types of meaningful land divisions such as the ecoregions described in Olson et al. (2001) or so called *pyromes*, either at global (Archibald et al., 2013) or regional (Trigo et al., 2016) scales, in order to further analyse the multiscalar nature of our approach.

- Extend the current study to perform predictions of BA for bimodal clusters. Although results for these regions have not been produced yet, the code required for that task is already in place.

- All the predictive models developed in this work use as predictors the set of observed climate indices described in Sec. 2.3. We plan to extend this catalog, not only by adding new climate teleconnection indices, but also by including other meaningful predictors which are more specific to fire, e.g. variables related to precipitation and soil moisture, as suggested in Archibald et al. (2013).

- In a next phase of the project, we will plan to assess the suitability of replacing observed by simulated teleconnection climate indices, as given by a set of state-of-the-art seasonal forecasting models, which can provide predictions of different atmospheric and oceanic variables (including the SST and SLP) for up to one year into the future. This would give us the chance to test the feasibility of developing an operational early warning system for the detection of high/low fire activity.

5.3 Reproducibility of Results

An additional effort has been undertaken in order to ensure the reproducibility of all the results presented in this work, adopting as far as possible the FAIR principles for scientific data management (Wilkinson et al., 2016). As a result, all the necessary data and code to reproduce the results are available in an open GitHub repository, including also several notebooks that serve as an aid in the full reproducibility of the results and their scrutiny, and constitute an extensive supplementary material to this study that is referred throughout the text.

All the resources needed to reproduce the results presented in this document are publicly available in *GitHub*: `https://github.com/MarcosVM98/TFM`. In particular, we have created a series of Jupyter notebooks (in *notebooks* directory), which build on the data and functions allocated in *data* and *scripts* directories respectively, and allow to easily follow all the analysis performed. Moreover, the code used for directly obtaining the figures that appear in this document (placed in *Figures* directory) is in *Figures* notebook. Finally, all the project was developed in the free *R* language.

5.4 Acknowledgements

The Global Burned Area data used in this study were downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store (`https://cds.climate.copernicus.eu`).

All the climate indices used in the empirical fire-climate models developed in this study have been downloaded from the NOAA's Climate Prediction Center (CPC, `https://psl.noaa.gov/data/climateindices/list/`).

Bibliography

Ameijeiras-Alonso, J., Benali, A., Crujeiras, R. M., Rodríguez-Casal, A., and Pereira, J. M. C. (2019). Fire seasonality identification with multimodality tests. *The Annals of Applied Statistics*, 13(4):2120–2139. URL `https://projecteuclid.org/euclid.aoas/1574910038`.

Archibald, S., Lehmann, C. E. R., Gómez-Dans, J. L., and Bradstock, R. A. (2013). Defining pyromes and global syndromes of fire regimes. *Proceedings of the National Academy of Sciences*, 110(16):6442–6447. URL `http://www.pnas.org/content/110/16/6442`.

Barnston, A. G. and Livezey, R. E. (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly weather review*, 115(6):1083–1126.

Bedia, J., Golding, N., Casanueva, A., Iturbide, M., Buontempo, C., and Gutiérrez, J. M. (2018). Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. *Climate Services*, 9:101–110.

Bedia, J., Herrera, S., Gutierrez, J., Benali, A., Brands, S., Mota, B., and Moreno, J. (2015). Global patterns in the sensitivity of burned area to fire-weather: implications for climate change. *Agricultural and Forest Meteorology*, 214–215:369–379.

Bedia, J., Herrera, S., and Gutierrez, J. M. (2014). Assessing the predictability of fire occurrence and area burned across phytoclimatic regions in Spain. *Natural Hazards and Earth System Science*, 14(1):53–66. URL `http://www.nat-hazards-earth-syst-sci.net/14/53/2014/`.

Benali, A., Mota, B., Carvalhais, N., Oom, D., Miller, L. M., Campagnolo, M. L., and Pereira, J. M. C. (2017). Bimodal fire regimes unveil a global-scale anthropogenic fingerprint: Benali et al. *Global Ecology and Biogeography*, 26(7):799–811. URL https://onlinelibrary.wiley.com/doi/10.1111/geb.12586.

Bond, W. J., Woodward, F. I., and Midgley, G. F. (2004). The global distribution of ecosystems in a world without fire. *New Phytologist*, 165(2):525–538. URL http://doi.wiley.com/10.1111/j.1469-8137.2004.01252.x.

Boschetti, L. and Roy, D. P. (2008). Defining a fire year for reporting and analysis of global interannual fire variability. *Journal of Geophysical Research: Biogeosciences*, 113(G3):G03020. URL http://onlinelibrary.wiley.com/doi/10.1029/2008JG000686/abstract.

Bowman, D. M. J. S., Balch, J., Artaxo, P., Bond, W. J., Cochrane, M. A., D'Antonio, C. M., DeFries, R., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Mack, M., Moritz, M. A., Pyne, S., Roos, C. I., Scott, A. C., Sodhi, N. S., and Swetnam, T. W. (2011). The human dimension of fire regimes on Earth: The human dimension of fire regimes on Earth. *Journal of Biogeography*, 38(12):2223–2236. URL http://doi.wiley.com/10.1111/j.1365-2699.2011.02595.x.

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., Swetnam, T. W., van der Werf, G. R., and Pyne, S. J. (2009). Fire in the Earth System. *Science*, 324(5926):481–484.

Brands, S., Manzanas, R., Gutiérrez, J. M., and Cohen, J. (2012). Seasonal predictability of wintertime precipitation in europe using the snow advance index. *Journal of Climate*, 25(12):4023–4028.

Bueh, C. and Nakamura, H. (2007). Scandinavian pattern and its climatic impact. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 133(629):2117–2131.

Chen, Y., Morton, D. C., Andela, N., Giglio, L., and Randerson, J. T. (2016). How much global burned area can be forecast on seasonal time scales using sea surface temperatures? *Environmental Research Letters*, 11(4):045001. URL http://stacks.iop.org/1748-9326/11/i=4/a=045001.

Chuvieco, E., Giglio, L., and Justice, C. (2008). Global characterization of fire activity: toward defining fire regimes from Earth observation data. *Global Change Biology*, 14(7):1488–1502. URL http://doi.wiley.com/10.1111/j.1365-2486.2008.01585.x.

Esbensen, S. K. (1984). A comparison of intermonthly and interannual teleconnections in the 700 mb geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, 112(10):2016–2032.

Fraley, C. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578–588. URL https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/41.8.578.

Frías, M. D., Iturbide, M., Manzanas, R., Bedia, J., Fernández, J., Herrera, S., Cofiño, A. S., and Gutiérrez, J. M. (2018). An R package to visualize and communicate uncertainty in seasonal climate prediction. *Environmental Modelling & Software*, 99:101–110. URL https://www.sciencedirect.com/science/article/pii/S1364815217305157.

Giglio, L., Randerson, J. T., and van der Werf, G. R. (2013). Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4): ANALYSIS OF BURNED AREA. *Journal of Geophysical Research: Biogeosciences*, 118(1):317–328. URL http://doi.wiley.com/10.1002/jgrg.20042.

Hantson, S., Kelley, D. I., Arneth, A., Harrison, S. P., Archibald, S., Bachelet, D., Forrest, M., Hickler, T., Lasslop, G., Li, F., Mangeon, S., Melton, J. R., Nieradzik, L., Rabin, S. S., Prentice, I. C., Sheehan, T., Sitch, S., Teckentrup, L., Voulgarakis, A., and Yue, C. (2020). Quantitative assessment of fire and vegetation properties in simulations with fire-enabled vegetation models from the Fire Model Intercomparison Project. *Geoscientific Model Development*, 13(7):3299–3318. URL https://gmd.copernicus.org/articles/13/3299/2020/.

Hurrell, J. W. (1995). Decadal trends in the north atlantic oscillation: Regional temperatures and precipitation. *Science*, 269(5224):676–679.

Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M., Manzanas, R., San-Martín, D., Cimadevilla, E., Cofiño, A., and Gutiérrez, J. (2019). The R-based climate4R open framework for reproducible climate data access and post-processing. *Environmental Modelling & Software*, 111:42–54. URL https://linkinghub.elsevier.com/retrieve/pii/S1364815218303049.

Jolly, W. M., Cochrane, M. A., Freeborn, P. H., Holden, Z. A., Brown, T. J., Williamson, G. J., and Bowman, D. M. J. S. (2015). Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6:7537. URL http://www.nature.com/doifinder/10.1038/ncomms8537.

Kelley, D. I., Prentice, I. C., Harrison, S. P., Wang, H., Simard, M., Fisher, J. B., and Willis, K. O. (2013). A comprehensive benchmarking system for evaluating global vegetation models. *Biogeosciences*, 10(5):3313–3340. URL https://bg.copernicus.org/articles/10/3313/2013/.

Kim, Y., Kim, K.-Y., and Jhun, J.-G. (2013). Seasonal evolution mechanism of the east asian winter monsoon and its interannual variability. *Climate dynamics*, 41(5-6):1213–1228.

Krawchuk, M. A. and Moritz, M. A. (2010). Constraints on global fire activity vary across a resource gradient. *Ecology*, 92(1):121–132. URL http://www.esajournals.org/doi/abs/10.1890/09-1843.1.

Krichak, S. O., Breitgand, J. S., Gualdi, S., and Feldstein, S. B. (2014). Teleconnection–extreme precipitation relationships over the mediterranean region. *Theoretical and applied climatology*, 117(3):679–692.

Le Page, Y., Oom, D., Silva, J. M. N., Jönsson, P., and Pereira, J. M. C. (2010). Seasonality of vegetation fires as modified by human action: observing the deviation from eco-climatic fire regimes. *Global Ecology and Biogeography*, 19(4):575–588. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1466-8238.2010.00525.x/abstract.

Lobell, D. B., Cahill, K. N., and Field, C. B. (2007). Historical effects of temperature and precipitation on California crop yields. *Climatic Change*, 81(2):187–203. URL http://link.springer.com/article/10.1007/s10584-006-9141-3.

Magi, B. I., Rabin, S., Shevliakova, E., and Pacala, S. (2012). Separating agricultural and non-agricultural fire seasonality at regional scales. *Biogeosciences*, 9(8):3003–3012. URL https://bg.copernicus.org/articles/9/3003/2012/.

Manzanas, R., Frías, M., Cofiño, A., and Gutiérrez, J. M. (2014). Validation of 40 year multimodel seasonal precipitation forecasts: The role of enso on the global skill. *Journal of Geophysical Research: Atmospheres*, 119(4):1708–1719.

Manzanas, R. and Gutiérrez, J. M. (2019). Process-conditioned bias correction for seasonal forecasting: a case-study with enso in peru. *Climate Dynamics*, 52(3):1673–1683.

Marcos, R., Turco, M., Bedia, J., Llasat, M. C., and Provenzale, A. (2015). Seasonal predictability of summer fires in a Mediterranean environment. *International Journal of Wildland Fire*, 24(8):1076–1084. URL http://www.publish.csiro.au/?paper=WF15079.

Murphy, K. P. (2012). *Machine Learning: A probabilistic Perspective*. The MIT Press.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., and Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, 51(11):933. URL https://academic.oup.com/bioscience/article/51/11/933-938/227116.

Pausas, J. G. and Ribeiro, E. (2013). The global fire-productivity relationship. *Global Ecology and Biogeography*, 22(6):728–736. URL http://doi.wiley.com/10.1111/geb.12043.

Pereira, J. M. C., Oom, D., Pereira, P., Turkman, A. A., and Turkman, K. F. (2015). Religious Affiliation Modulates Weekly Cycles of Cropland Burning in Sub-Saharan Africa. *PLOS ONE*, 10(9):e0139189. URL http://dx.plos.org/10.1371/journal.pone.0139189.

Preisler, H. K. and Westerling, A. L. (2007). Statistical Model for Forecasting Monthly Large Wildfire Events in Western United States. *Journal of*

Applied Meteorology and Climatology, 46(7):1020–1030. URL http://journals.ametsoc.org/doi/abs/10.1175/JAM2513.1.

Rodrigues, M., Peña-Angulo, D., Russo, A., Zúñiga-Antón, M., and Cardil, A. (2021). Do climate teleconnections modulate wildfire-prone conditions over the Iberian Peninsula? *Environmental Research Letters*, 16(4):044050. URL https://iopscience.iop.org/article/10.1088/1748-9326/abe25d.

Rogers, J. C. (1997). North atlantic storm track variability and its association to the north atlantic oscillation and climate variability of northern europe. *Journal of Climate*, 10(7):1635–1647.

Scrucca, L., Fop, M., Murphy, T., Brendan, and Raftery, A., E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1):289. URL https://journal.r-project.org/archive/2016/RJ-2016-021/index.html.

Taschetto, A. S., Ummenhofer, C. C., Stuecker, M. F., Dommenget, D., Ashok, K., Rodrigues, R. R., and Yeh, S.-W. (2020). Enso atmospheric teleconnections. *El Niño Southern Oscillation in a Changing Climate*, pp. 309–335.

Trenberth, K. E. (1984). Signal versus noise in the southern oscillation. *Monthly Weather Review*, 112(2):326–332.

Trigo, R. M., Sousa, P. M., Pereira, M. G., Rasilla, D., and Gouveia, C. M. (2016). Modelling wildfire activity in iberia with different atmospheric circulation weather types. *International Journal of Climatology*, 36(7):2761–2778. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3749.

Turco, M., Jerez, S., Doblas-Reyes, F. J., AghaKouchak, A., Llasat, M. C., and Provenzale, A. (2018). Skilful forecasting of global fire activity using seasonal climate predictions. *Nature Communications*, 9(1):2718. URL http://www.nature.com/articles/s41467-018-05250-0.

Turco, M., Llasat, M.-C., von Hardenberg, J., and Provenzale, A. (2014). Climate change impacts on wildfires in a Mediterranean environment. *Climatic Change*, 125:369–380. URL http://link.springer.com/10.1007/s10584-014-1183-3.

Urbieta, I. R., Zavala, G., Bedia, J., Gutiérrez, J. M., Miguel-Ayanz, J. S., Camia, A., Keeley, J. E., and Moreno, J. M. (2015). Fire activity as a function of fire–weather seasonal severity and antecedent climate across spatial scales in southern Europe and Pacific western USA. *Environmental Research Letters*, 10(11):114013. URL `http://stacks.iop.org/1748-9326/10/i=11/a=114013?key=crossref.464a9e70dbebd8a2f723e7dd8fdb3490`.

Van Loon, H. and Rogers, J. C. (1978). The seesaw in winter temperatures between greenland and northern europe. part i: General description. *Monthly Weather Review*, 106(3):296–310.

Wallace, J. M. and Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly weather review*, 109(4):784–812.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018. URL `http://www.nature.com/articles/sdata201618`.

Zveryaev, I. (2009). Interdecadal changes in the links between european precipitation and atmospheric circulation during boreal spring and fall. *Tellus A: Dynamic Meteorology and Oceanography*, 61(1):50–56.