Performance of Statistical and Machine Learning-Based Methods for Predicting Biogeographical Patterns of Fungal Productivity in Forest Ecosystems

Albert Morera (morera.marra@gmail.com)
UdL: Universitat de Lleida https://orcid.org/0000-0002-6777-169X
Juan Martínez de Aragón
Forest Science and Technology Centre of Catalonia
José Antonio Bonet
Escola Tècnica Superior d'Enginyeria Agrària de Lleida: Universitat de Lleida Escola Tecnica Superior d'Enginyeria Agraria
Jingjing Liang
Purdue University Department of Forestry and Natural Resources
Sergio de-Miguel
Escola Tècnica Superior d'Enginyeria Agrària de Lleida: Universitat de Lleida Escola Tecnica Superior d'Enginyeria Agraria

Research

Keywords: Artificial intelligence, Modelling, Biogeography, Climate, Forest, Fungi, Mushrooms

Posted Date: December 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-122045/v1

Version of Record: A version of this preprint was published on March 15th, 2021. See the published version at https://doi.org/10.1186/s40663-021-00297-w.

Title: Performance of statistical and machine learning-based methods for predicting biogeographical patterns of fungal productivity in forest ecosystems

Authors: Albert Morera[1,2*], Juan Martínez de Aragón[3], José Antonio Bonet[1,2], Jingjing Liang[4], Sergio de-Miguel[1,2]

Affiliations:

[1] Department of Crop and Forest Sciences, University of Lleida, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain

[2] Joint Research Unit CTFC – AGROTECNIO, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain

[3] Forest Science and Technology Centre of Catalonia, Ctra. Sant Llorenç de Morunys km 2, 25280 Solsona, Spain

[4] Forest Advanced Computing and Artificial Intelligence Laboratory, Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA

* Corresponding author's email: morera.marra@gmail.com

Abstract

Background: The prediction of biogeographical patterns from a large number of driving factors with complex interactions, correlations and non-linear dependences require advanced analytical methods and modelling tools. This study compares different statistical and machine learning models for predicting fungal productivity biogeographical patterns as a case study for the thorough assessment of the performance of alternative modelling approaches to provide accurate and ecologically-consistent predictions.

Methods: We evaluated and compared the performance of two statistical modelling techniques, namely, generalized linear mixed models and geographically weighted regression, and four machine learning models, namely, random forest, extreme gradient boosting, support vector machine and deep learning to predict fungal productivity. We used a systematic methodology based on substitution, random, spatial and climatic blocking combined with principal component analysis, together with an evaluation of the ecological consistency of spatially-explicit model predictions.

Results: Fungal productivity predictions were sensitive to the modelling approach and complexity. Moreover, the importance assigned to different predictors varied between machine learning modelling approaches. Decision tree-based models increased prediction accuracy by ~7% compared to other machine learning approaches and by more than 25% compared to statistical ones, and resulted in higher ecological consistence at the landscape level.

Conclusions: Whereas a large number of predictors are often used in machine learning algorithms, in this study we show that proper variable selection is crucial to create robust models for extrapolation in biophysically differentiated areas. When dealing with spatial-

temporal data in the analysis of biogeographical patterns, climatic blocking is postulated as a highly informative technique to be used in cross-validation to assess the prediction error over larger scales. Random forest was the best approach for prediction both in sampling-like environments as well as in extrapolation beyond the spatial and climatic range of the modelling data.

Keywords

Artificial intelligence; Modelling; Biogeography; Climate; Forest; Fungi, Mushrooms

1 Background

Understanding the biogeographical patterns of organisms in natural ecosystems and predicting their distribution is a fundamental challenge in environmental sciences (Ehrlén and Morris, 2015). This entails a deep understanding of their distribution across space and time underpinning ecological mechanisms, which becomes increasingly complex with an increasing amount of factors driving these patterns and the possible interactions and nonlinear dependencies between them (Dixon et al., 1999; Ye et al., 2015). Such complex interrelationships require advanced data analytic methods and modelling tools to yield realistic predictions of natural ecosystem attributes and processes.

Statistical methods traditionally used for this purpose aim at accounting for several elements that govern these natural mechanisms, trying to reach a parsimonious and robust understanding of ecological patterns (Wood and Thomas, 1999). However, since conventional parametric approaches may over-simplify nonlinear relationships between variables and over- or under-estimate the influence of some drivers, conventional parametric approaches may result in poor predictions and/or descriptions of reality (Ye et al., 2015), especially for the analyses of large databases. To overcome potential limitations of classic statistical approaches in big data analysis, the increased computing power has led to recent considerable growth in the use of analytical methods based on artificial intelligence such as machine learning (Christin et al., 2019).

Machine learning algorithms are increasingly being used in species distribution and ecological niche modelling (Prasad et al., 2006; Culter et al., 2007; Hannemann et al., 2015; Liang et al. 2016; Prasad, 2018; Gobeyn et al., 2019), forest resources (Stojanova et al., 2010; Görgens et al., 2015) and climate change studies (Thuille, 2003; Bastin et al., 2019), among others. To determine to what extent these "new" methodologies can contribute to improving our understanding and prediction capacity within the field of

environmental sciences, comparative studies are required between those models that have been used historically and those fed by artificial intelligence algorithms (Özçelik et al., 2013; Diamantopoulou et al., 2015; Hill et al., 2016; Bonete et al., 2020). Yet, many machine learning algorithms have been developed in recent years, and each of them may be more or less appropriate depending on the specific tasks and research objectives (Thessen, 2016). This highlights the need for systematic studies allowing for discerning the most suitable methodology according to a given research objective and data. Although several studies have analysed the performance of different analytical approaches (Hill et al., 2016; Bonete et al., 2020; Mayfield et al., 2020), existing ecological research addressing systematic assessments and comparisons of alternative modelling and predictive methods is scarce, making it difficult to provide clear methodological recommendations about the suitability of different approaches. Besides, in the field of environmental sciences, often, extrapolations in biophysically differentiated areas are required, which makes it necessary to take even more into account the data spatial dependencies. Due to data spatial autocorrelation, random cross-validations lead to over-optimistic error estimates (Bahn and McGill, 2013; Micheletti et al., 2014; Juel et al., 2015; Gasch et al., 2015; Roberts et al., 2017; Meyer et al., 2018; Meyer et al., 2019a), which makes it necessary to use proper, complementary validation methods such as spatial cross-validation (Le Rest et al., 2014; Pohjankukka et al., 2017; Roberts et al., 2017; Meyer et al., 2018; Valavi et al., 2018). Moreover, spatial dependencies in the data can lead to a misinterpretation of some predictors outside the sampling range (Meyer et al., 2018, 2019a).

Biogeographical patterns of fungal dynamics over large scales are a highly relevant question in ecology given the key role of fungi in forest ecosystems (Stokland et al., 2012; Mohan et al., 2014), especially in fungi-tree symbiosis. However, due to their great

117 diversity and differential ecological requirements (Glassman et al., 2017), as well as the
118 difficulty of monitoring their dynamics and the large array of potential drivers (Büntgen
119 et al., 2013), little is known of fungal dynamics over large scales. The prediction of
120 biogeographical patterns of fungal dynamics requires large fungal datasets with a correct
121 taxonomic identification of the specimens and a consistent sampling methodology across
122 space and time to avoid sample bias (Hao et al., 2020).

123 In particular, the spatially-explicit prediction of fungal productivity, i.e. mushroom
124 fruiting patterns, is a key feature of fungal dynamics inasmuch as it is tightly related to
125 the supply of multiple provisioning, regulating and cultural ecosystem services (Boa,
126 2004). However, the high correlation between mushroom production and meteorological
127 conditions among other drivers (Taye et al., 2016; Alday et al., 2017; Collado et al. 2019)
128 makes the prediction of mushroom production challenging, especially in Mediterranean
129 ecosystems where there is a high inter-annual variability of climatic conditions. The long
130 period of potential fruiting of different mushroom species, as a result of their adaptation
131 to the recurrent climatic patterns of a dry summer followed by a wet autumn (Barnard et
132 al., 2015), makes mushroom yields dependent on a large number of variables.
133 Precipitation and temperatures on a weekly scale can be the factors that lengthen, shorten
134 or shift the fruiting period (Gange et al., 2007; Kauserud et al., 2008; Kauserud et al.,
135 2009; Büntgen et al., 2012), and also those that modulate mushroom production to a
136 higher degree (Karavani et al., 2018). The large number of variables involved and their
137 presumed interactions may often yield a misconception that fungal productivity is highly
138 stochastic or very difficult to predict. Previous research to estimate mushroom
139 productivity over large scales has been mainly based on mixed-effects modelling (de-
140 Miguel et al., 2014; Sánchez-González et al., 2019). Despite being a valid approach, it

may have certain limitations that are worth assessing in comparison with alternative methods that remain unexplored.

This study compares different statistical and machine learning models in estimating mushroom productivity at the landscape level, together with a systematic methodology to determine the best approach to predict mushroom production in forest ecosystems. Using climatic and biophysical data together with in situ fungal records collected weekly over more than 20 years on a hundred permanent plots, we developed spatially explicit, high-resolution continuous maps of mushroom productivity that were also used in the selection of the most suitable methods for predicting this ephemeral and important forest resource. We tested six most widely used predictive models to date, namely, generalized linear mixed-effects models (GLMM), and compared them with geographically weighted regression models (GWR), as well as with alternative state-of-the-art machine learning algorithms such as random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and deep learning (DL) models.

2 Methods

2.1 Study area and sampling plots

The study area was Catalonia region, northeastern Spain, in the western Mediterranean basin. The forest ecosystem types considered in this study were the main Mediterranean pine forest ecosystems that represent the majority of the forest area of the study region, namely, pure stands of *Pinus halepensis*, *P. sylvestris*, *P.pinaster*, *P.nigra* and *P.uncinata* and mixed stands of *P. halepensis* and *P. nigra*, and of *P. sylvestris* and *P. nigra*. We used a dataset that gathered information from 98 permanent monitoring plots for fungal dynamics sampled on a weekly basis during the main mushroom fruiting period, between August and the end of December and from 1997 to 2019. The plots were distributed

randomly and proportionally to the relative surface of the different pine forest ecosystems (Bonet et al., 2010) (Figure S1). Data were aggregated to an annual basis to create predictive models to estimate annual mushroom productivity. More information about the sampling methods and data can be found in Bonet et al. (2004), Martínez de Aragón et al. (2007) and Table 1.

Table 1. Summary of mushroom, climate and physical data of the 98 sampled plots.

	Min.	1st quart.	Median	Mean	3rd quart.	Max.
Mushroom yield (kg ha^{-1} $year^{-1}$)	0.00	9.14	47.62	104.46	138.97	1345.78
Slope (degrees)	3.00	12.00	18.00	18.07	23.00	37.00
Aspect (degrees)	4.00	58.00	179.00	166.50	282.00	360.00
August precipitation (mm)	0.00	14.56	29.07	42.10	62.10	170.89
September precipitation (mm)	0.21	37.59	54.94	59.01	77.51	202.79
October precipitation (mm)	4.62	32.50	56.08	73.46	90.13	254.23
August mean temperature (ºC)	14.62	20.03	22.71	22.05	24.20	27.31
September mean temperature (ºC)	9.47	16.10	18.39	17.86	20.02	23.74
October mean temperature (ºC)	7.24	12.36	14.15	13.81	15.58	19.19
August max. temperature (ºC)	17.93	25.08	28.34	28.12	30.27	39.26
September max. temperature (ºC)	13.61	21.04	23.48	23.43	25.84	33.24
October max. temperature (ºC)	10.98	16.77	18.76	19.03	20.84	29.05
August min. temperature (ºC)	0.93	11.95	14.36	13.27	15.87	19.04
September min. temperature (ºC)	-1.19	8.50	10.77	9.86	12.33	17.29
October min. temperature (ºC)	-6.80	4.95	7.46	6.24	8.83	13.50

2.2 Climate and biophysical data

Meteorological data for each sampling plot was obtained from the interpolation and altitudinal correction of daily weather of 201 meteorological stations from the Catalan Meteorological Service and the Spanish Meteorological Agency. Interpolation was conducted with "meteoland" R package (De Cáceres et al., 2018) that uses a modification of the DAYMET methodology (Thornton et al., 1997; Thornton and Running, 1999). Likewise, to determine the typical climatic conditions across the whole study region, we used the mean of the interpolated daily weather variables for the period between 1991 and 2016 with 1-km resolution. We computed the accumulated monthly rainfall from August to October and the mean, maximum and minimum monthly temperatures for the same period, coinciding with the main mushroom fruiting period.

The total area covered by the different pine forest ecosystems was retrieved from the CORINE habitats map (Commission of the European Community, 1991). The biophysical variables such as elevation, slope, aspect and stand basal area were obtained at 20-m resolution from the first cover of the LIDARCAT Project (http://territori.gencat.cat/es/detalls/Article/Mapes_variables_biofisiques_arbrat) based on different LiDAR flights between 2008 and 2011 with a point density of 0.5 points/m^2.

2.3 Analytical methods

We used and compared six different analytical methods to predict annual mushroom productivity. Two analytical approaches were based on statistical methods, namely, generalized linear mixed-effects models (GLMM) and geographically weighted regression (GWR), whereas the other four analytical methods were based on alternative machine learning approaches, namely, random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and deep learning (DL).

197 2.3.1 Statistical modelling

198 We used a two-stage modelling approach to take into account the high frequency of
199 "zero" production values in many sample plots over time (Hamilton and Brickell, 1983;
200 de-Miguel et al., 2014; Karavani et al., 2018; Collado et al., 2018). The high occurrence
201 of these values arise from the small size of the plots and the stochastic nature of
202 mushroom emergence (de-Miguel et al., 2014).

203 The first stage determines the probability of mushroom emergence, according to binomial
204 data of presence/absence, using a logistic regression $\pi(X)=E(Y|X)$ and a logit link
205 function to represent the conditioned mean of Y given X (Eq. 1 and 2).

206 $$\pi(x_k) = E(Y|x_k) = \frac{1}{1+e^{-g(x_k)}} \quad (1)$$

207 $$g(x_k) = \ln\left(\frac{\pi(x_k)}{1-\pi(x_k)}\right) = \beta_0 + \beta_k x_k \quad (2)$$

208 where $\pi(x_k)$ is the probability of mushroom occurrence, $g(x_k)$ the logit transformation of
209 $\pi(x_k)$, Y is the dependent variable (mushroom presence/absence), x_k is k^{th} independent
210 variable, β_0 is the intercept parameter and β_k is the regression coefficient for k^{th}
211 independent variable.

212 The second stage was based on the modelling of the production of non-zero production
213 values at logarithmic scale using linear regression $y=E(\log(Y)|X) + \varepsilon$. Logarithmic
214 transformation allows to limit the production range in the interval $[0, \infty)$, depending on
215 the values of X (Eq. 3). The proportional bias of the logarithmic regression was corrected
216 with the Snowdon's bias correction factor (Snowdon, 1991) based on the ratio of the
217 arithmetic sample mean and the mean of the back-transformed predicted values from the
218 regression (Eq. 4):

219 $$\ln(prod) = \beta_0 + \beta_k x_k + \varepsilon \quad (3)$$

$$CF = Y/\hat{y} \quad (4)$$

where ln(*prod*) represents the non-zero production of mushrooms at logarithmic scale, β_0 is the intercept parameter, β_k the regression coefficient for k^{th} independent variable, ε the random error of the deviation of the observations from the conditioned mean of ln(*Y*) and Y/=ŷ is the ratio between the mean of observed and the mean of predicted values of the sapling units.

Finally, the total production of mushrooms was obtained from the product of the probability of appearance and the conditioned production of non-zero values (Eq. 5).

$$yield = \pi(x_k)\, e^{\ln(yield)}\, CF \quad (5)$$

where $\pi(x_k)$ is the probability of mushroom occurrence, ln(*yield*) represents the production of mushrooms at logarithmic scale and *CF* is bias correction factor.

2.3.1.1 Generalized Linear Mixed Models

Due to mushrooms sampling methodology, where annually data was taken from a network of permanent plots, we used Generalized Linear Mixed Models GLMM (de-Miguel et al., 2014; Karavani et al., 2018; Collado et al., 2018). This method can consider the spatial and temporal autocorrelation among observations (Pinheiro and Bates, 2000) adding random effects to segment the data into different groups according to year and plot. In the proposed mixed-effects models only random effects on model interception were considered. All the models were fitted using the "glmer" function from the "lme4" R package (Bates et al., 2014).

2.3.1.2 Geographically Weighted Regression

GWR is a non-stationary modelling technique that describes the spatially varying relationships between the dependent variable and the explanatory variables (Páez and

Wheeler, 2009). Coefficients of a GWR-based model are given by the spatial location of data and can be estimated for any new location. This means that given a grid, the estimated coefficients for each point in space vary continuously as a function of the spatial heterogeneity of the relationships.

Coefficients for each regression point were calibrated using the data around itself. Due to the annual sampling methodology and the geographical distribution of plots, some plots were grouped denser in some areas and less dense in others. Consequently, we used an adaptive window according to the spatial density of our plots (Georganos et al., 2017). Occurrence and conditional production models, as well as the optimal data value for adjusting the adaptive window, were obtained from the "ggwr" and "gwr.set" functions, respectively, of the R package "spgwr" (Bivand el al., 2014).

2.3.2 Machine learning modelling

2.3.2.1 Decision trees-based models (random forest and extreme gradient boosting)

RF algorithm (Breiman, 2001) is based on an ensemble learning method that detects general patterns present in observed data using a defined number of decision trees. Trees are grown independently of each other, achieving less correlation between them. This is achieved by selecting a random subset of the predictors, in each node, to be chosen as the best candidate for use in each of the divisions (Breiman, 2001).

The hyperparameters that we tuned in the RF models were the number of trees and the optimal number of predictor candidates to be chosen in each node. The first one was tuned by training a sequence of models with 1 tree and up to 500, other things being equal, to locate the minimum number of trees required to minimize the prediction error. Once this value was set, we evaluated the optimal number of predictor candidates to be chosen in each node in a similar way, picking the one that minimizes the error while growing less

267 correlated forests. In both cases, cross-validation was used (randomly splitting 95% of
268 the data for training and the remaining 5% for validation) and 20 models were trained for
269 each combination of hyperparameters (Table S3). Hyperparameters tuning and model
270 training were done with the R-package "RandomForest" (Liaw and Wiener, 2002).

271 Unlike RF, the technique of gradient boosting (Friedman, 1999) does not create each
272 decision tree independent of each other, but rather the trees are built sequentially. Thus,
273 each constructed tree uses the information on the previously cultivated trees (Jamet et al.,
274 2013). XGB is a specific implementation of the boosting method that allows finding the
275 best tree by minimizing the loss function using a minimum amount of resources (Chen
276 and Guestrin, 2016). To minimize model overfitting, XGB is basically based on two
277 techniques: shrinkage and column subsampling. This means that, in each iteration, a
278 fraction of the observed data is used to grow the next tree, using the weighted information
279 from each of the previous trees.

280 As in RF-hyperparameters tuning, we used cross-validation by training 20 models for
281 each hyperparameters combination. We tuned the number of iterations, in a range
282 between 1 and 100, according to the learning rate (eta) to improve the prediction accuracy
283 and optimize the XGB algorithm. Next, we tuned the depth of the model trees and finally,
284 since this depends on the others, the gamma regularization parameter to avoid model
285 overfitting (Table S4). The R-package "xgboost" (Chen and He, 2020) was used here to
286 train the XGB models and tune their hyperparameters.

287 2.3.2.2 Support Vector Machine

288 SVM (Cortes and Vapnik, 1995) was created to solve classification problems by adjusting
289 a hyperplane that delimited the different categories while minimizing the distance
290 between the hyperplane and the closest points. Although its main purpose was to be used

in classification problems, later updates have made this algorithm applicable to regression tasks, using the same principles with which it was conceived. In the case of regression, a tolerance margin is established, corresponding to the support vectors, and an appropriate hyperplane is defined to adjust the data.

A radial-based kernel function was used both to train the SVM models and to select their optimal hyperparameters. The cost and gamma hyperparameters were tuned to each other based on a cross validation of the predictions made by 20 models, with values between 1 and 50 and between 0.1 and 0.9, respectively (Tale S5). SVM models and their hyper-parameter tuning were done using the R-package "e1071" (Meyer et al., 2019).

2.3.2.3 Deep Learning

Machine learning includes those algorithms that generate models from patterns in a data set, making it interesting for the analysis of complex and non-linear data (Olden et al., 2008). While other machine learning algorithms find patterns among the variables used to train the models, DL is able to detect and extract undescribed characteristics directly among the highly dimensioned data (Christin et al., 2019). These characteristics have led to an increase in the use of these algorithms in recent years in ecology (Christin et al., 2019).

As in the tuning of the best hyperparameters for each of the others machine learning model, we proposed different candidate structures, given by the following set of hyperparameters, for the DL models network. We trained models with 1, 2 and 3 hidden layers with a "relu" regularization function and a final layer with a single node and the same regularization function. For each structure, we used hidden layers with 8, 16, 32, 64 and 128 nodes in each one. To reduce overfitting, we randomly dropped out (setting to zero) 50% of output features of the layer during training. We evaluated each structure

315 from the cross-validation of the predictions of 20 models to minimize the prediction error
316 (Table S6). We used the R-package "keras" (Allaire and Chollet, 2019) to train the final
317 DL models and the different proposed structures.

318 2.3.3 Model and variable selection and evaluation

319 Statistical model evaluation and variable selection was based on the current knowledge
320 of forests and mushroom ecology, the statistical significance of model parameters
321 ($p<0.05$ or $t>|1.96|$), the variance inflation factor (VIF) to quantify the severity of
322 multicollinearity and the parsimony principle. To check the sensitivity/specificity of the
323 binomial classification models we used Receiver Operator Characteristic (ROC) curves,
324 using the R package “ROCR” (Sing et al., 2005).

325 To assess whether GWR improved GLMM due to the non-stationary nature of data and
326 to avoid introducing an improvement that was not attributed to the type of modelling, the
327 same explanatory variables as in GLMM were used. To test the non-stationarity of the
328 independent variables of GWR models, the local parameters were compared with global
329 GLMM coefficients. The probability of incorporating non-stationary variables increases
330 if the estimated coefficient of the variable in GLMM (± standard error) is outside the 1st
331 and 3rd quartile of the GWR model coefficient (Propastin, 2009).

332 For each of the four machine learning algorithms, two types of models were adjusted. The
333 first ones were trained using a total of 15 biophysical variables (Table 1), while the second
334 ones were trained using a subset of them. This subset was determined from the same 5
335 variables used in the statistical models, including climate predictors only. This allowed
336 us to assess separately the prediction accuracy due to the analytical method, and the
337 prediction accuracy due to differences between predictors or the number of explanatory
338 variables.

Due to the randomness in the training stage of RF and XGB algorithms, and for consistency with the methodology used for SVM and DL models, 50 models were trained from 50 bootstrapping iterations, selecting randomly 95% of the data.

In machine learning-based approaches, the relationship between predictors and mushroom productivity was assessed based on partial dependence plots (PDPs), a low-dimensional graphical rendering between variable pairs, in order to determine whether this relationship lacked ecological sense (within a set of specific predictors, and in this particular case, corresponding to the training data set).

2.3.4 Evaluation of the predictive performance and mapping

Since the main purpose of this study was to develop models to accurately predict mushroom productivity, this entails the evaluation of the predictive performance of the resulting models also outside of the range of the training (or fitting) region. With the aim of determining the similarities between sampled plots and the territory where it was used to estimate the mushroom yield, a principal component analysis (PCA) based on the models' predictors was used. In addition, to evaluate and compare the predictive accuracy of the models, different cross-validation strategies were used (Roberts et al., 2017). This validation was done by comparing the determination coefficient (R2), root mean squared error (RMSE) and mean absolute error (MAE) of observed (using 95% of the data) against predicted (i.e. the remaining 5% of the data) values of each model using 50 bootstrapping iterations. To inspect in depth, the source of the errors in the models, the mean square deviation (MSD) was estimated and also decomposed in additive components, namely, squared bias (SB), nonunity slope (NU) and lack of correlation (LC) (Gauch et al., 2003). The alternative cross-validation strategies utilized were:

i. Substitution: the testing data was part of the same training data

ii. Random: the training and testing data set were selected randomly

iii. Spatial blocking: to avoid underestimating the prediction error due to the spatial structure of the data the training and testing data were strategically and non-randomly selected (Roberts et al., 2017). The two data sets were selected in different geographical areas, selecting a randomly plot and blocking the closest plots as testing data.

iv. Climate blocking: since plots were spatially distributed among different bioclimatic areas, we conducted an environmental blocking (Valavi et al., 2018) considering the differences in climatic conditions between plots. Similar to spatial blocking (Roberts et al., 2017), we used the two-dimensional space described by the first two PCA components (based on the sets of five and twelve climatic model predictors, respectively) to select the training and testing datasets from different climatic areas. In this 2-D space, one plot was randomly selected and blocked as testing data together with a giver number of neighbouring plots.

A 10% and 20% blocking was also considered to determine the error in climatic and geographical areas more different from the sampled data (as a comparison it was also done in the random blocking), providing a more differentiated dataset of training and testing.

To generate the landscape-level mushroom productivity maps based on RF, XGB, SVM and DL, we used the mean estimation of the 50 trained models. Otherwise, estimates from statistical models were the result of the empirical model predictions, fitted with 100% of data. These maps were constructed with a resolution of 1 km in accordance with the resolution of the climatic data. As a result of the estimates of the 50 models, spatially explicit relative standard error (SE) maps were also created for each methodology.

386 The final maps were evaluated on the basis of knowledge about fruiting patterns, and also
387 on whether they followed ecologically logical patterns (related to climatic conditions).
388 Therefore, we would expect smoothed estimates across the territory and conditioned by
389 the variations of those most important variables in each model.

390 **3 Results**

391 3.1 Relationships between dependent and explanatory variables

392 Statistical models showed a statistically significant and positive relationship of mushroom
393 productivity with rainfall in August, September and October (both in conditioned
394 production and occurrence models). On the other hand, conditioned production and
395 occurrence models also showed a statistically significant and negative relationship with
396 the maximum temperature of August and October, respectively (Tables S1 to S3). Yet,
397 the coefficients of GWR models varied according to geographical location (Table S3 and
398 Figure S2), describing certain non-stationarity in both precipitation and temperature.

399 Within GLMM models, PDPs showed an almost linear relationship between the amount
400 of precipitation between August and October and mushroom productivity in the model fit
401 data range. In contrast, GWR showed an accelerated growth in productivity by increasing
402 rainfall, which was accentuated in those locations with a higher precipitation regression
403 coefficient. Besides, and similarly in GLMM and GWR, the maximum temperature in
404 August showed a decelerated decrease in productivity by increasing temperature, while
405 the maximum temperature of October, even though it showed a negative relation, resulted
406 in little relevance in mushroom productivity for the range of values of the fitting data
407 (Figure 1).

408 Different machine learning models resulted in rather similar relationships between
409 variables although, due to the particularities of each algorithm, the patterns changed

slightly between approaches. In contrast to the relationships in GLMM and GWR models, some of the machine learning models did not show monotonically increasing or decreasing relationships between dependent and explanatory variables. This monotony was often broken at the extremes of the range of values of the predictor variables, where the amount of data to train the models was lower (Figure 1). Moreover, machine learning methods also showed differences in the importance assigned to different predictors. Thus, XGB identified some variables as very important compared to other predictors. Specifically, in the models trained with 15 variables, XGB showed a greater importance of precipitation in August, September and October, minimum temperature in October and aspect. In addition, precipitation of August resulted in having further greater importance in the models trained with 5 variables. Conversely, the importance detected by RF and SVM to the whole array of predictors was more homogeneous. RF showed a greater importance to the same variables as XGB, while the most important variables in SVM were precipitation in September and October, average temperature in August and September, and minimum temperature in August (Figure 2).

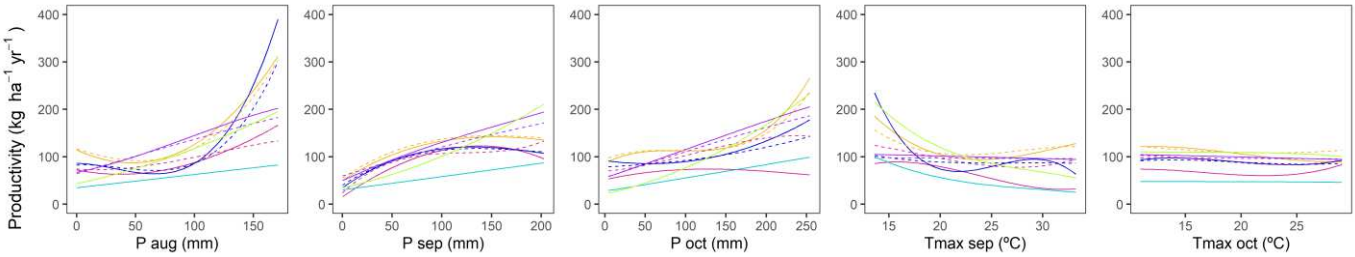


Figure 1. Relationship between annual mushroom productivity and August, September and October precipitation and maximum temperatures in September and October (these variables are the variables used in the statistical models and the five variables machine learning models). Yellow (random forest), blue (extreme gradient boosting), violet (support vector machine), purple (deep learning), light blue (generalized linear mixed models) and light green (geographically weighted regression) colours refer to the

different modelling techniques. Continuous line includes the models that use the five variables and the dashed line the machine learning models trained with 15.

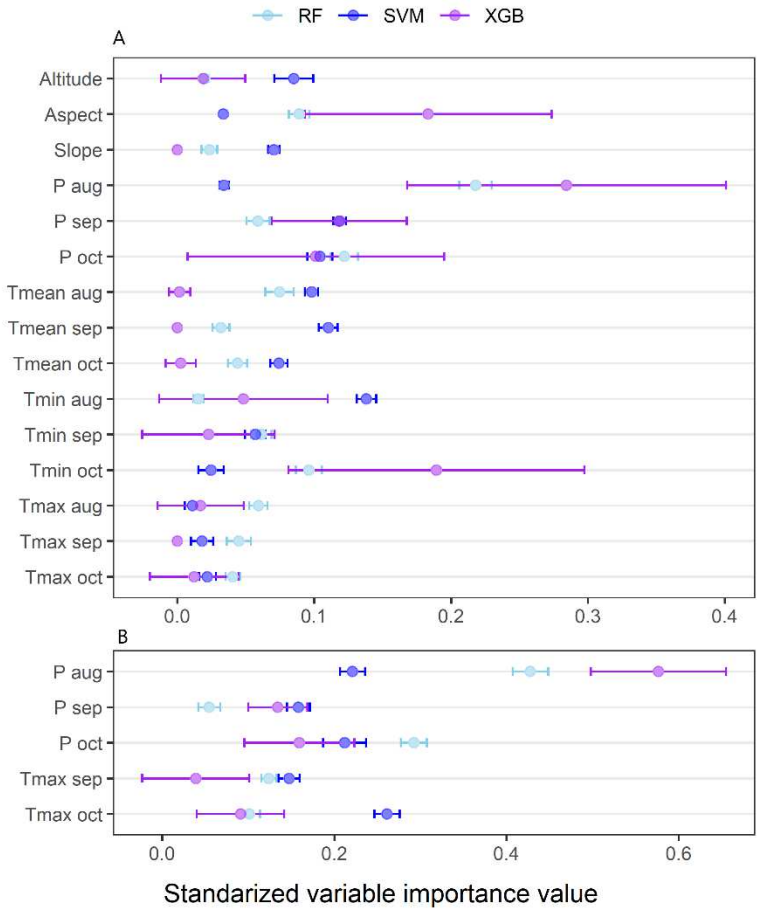
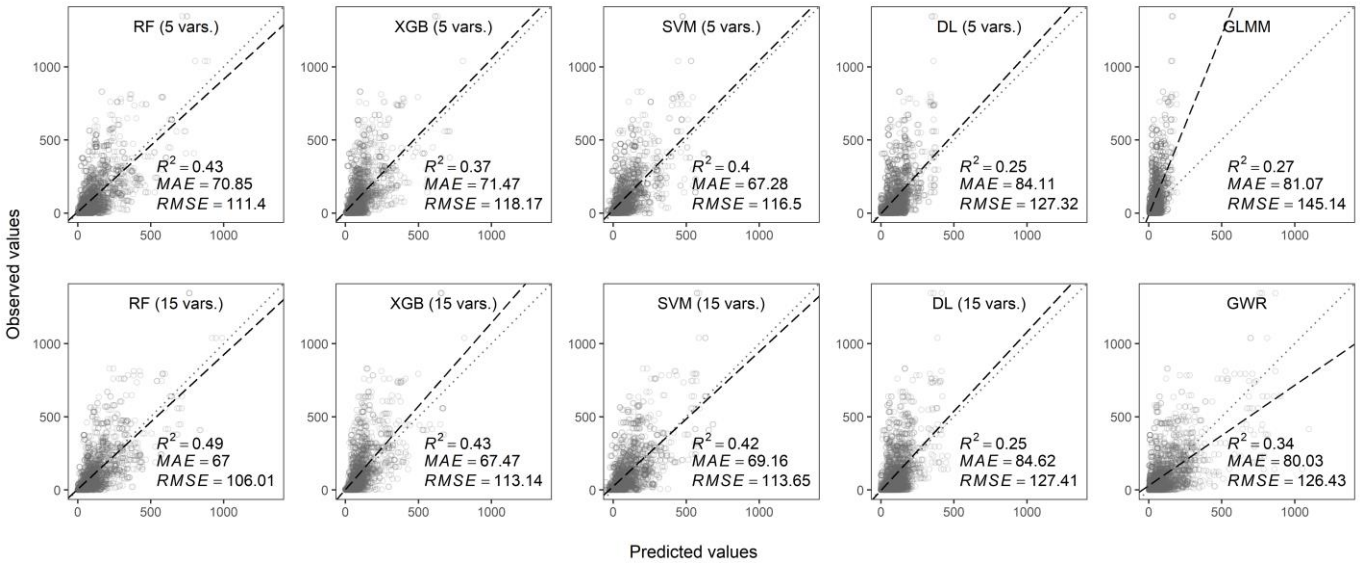


Figure 2. Standardized variable importance value used to train random forest (RF), extreme gradient boosting (XGB) and support vector machine (SVM) models with 15 (A) and 5 (B) variables. Variable importance values represent the contribution of each variable to predict the annual mushroom productivity.

GLMM and GWR fitted models and their coefficients are shown in the supplementary material in Table S1 to S3. Likewise, machine learning hyperparameters can be found in the supplementary material in Figure S3 to S6.

443 3.2 Predictive accuracy of different methods

444 In general, machine learning models showed better predictive accuracy compared to
445 statistical models. Above all, models based on decision trees, and more specifically RF,
446 stood out as the models with the lowest error. RF, XGB and SVM models trained with 15
447 variables showed lower errors than the same ones trained with five variables, while DL
448 models resulted in similar prediction error regardless the amount of predictors. Moreover,
449 RF trained with five variables resulted in lower or similar error compared to XGB and
450 SVM models trained with 15 variables (Figure 3 and Table S4).



451 Figure 3. Cross-validation of estimated total mushroom production (kg ha-1 year^{-1}) using
452 random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM)
453 and deep learning (DL) trained with 15 and 5 variables and generalized linear mixed
454 models (GLMM) and geographically weighted regression (GWR). Pointed line shows the
455 equality line and the dashed one the best fit between predicted and observed values. It
456 shows mean absolute error (MAE), root-mean-squared error (RMSE) and coefficient of
457 determination (R^2) of the relation for each technique.

458 The assessment of the predictive performance based on random blocking resulted in lower
459 accuracy compared to spatial and climatic blocking in RF, XGB and SVM models. This

trend was not so obvious in DL or statistical models. In climate blocking, the prediction error increased with increasing proportion of blocking data, denoting lower predictive accuracy of fungal productivity (Table S4) in areas with very different climatic conditions compared to the training data (Figure 4). Most machine learning models, based on finding patterns in supplied training data, showed very low errors when making predictions in a similar environment as the training data (substitution), i.e. approximately 50% and 75% lower error in RF and XGB, respectively, compared to random blocking. Overall, RF was the approach with lower error, being lower in models trained with 15 variables than those trained with five variables (105 and 114 respectively in a random blocking and 110 and 111 in a climate blocking). Compared to the other tested models (in terms of RMSE in random blocking), RF results in an improvement between 7 and 31% in the models trained with 15 variables and between 4 and 26% in the RF models trained with 5 variables (Table S4).

Although in all the analytical approaches the main source of the error was due to the variation between training and testing data (i.e. lack of correlation, LC), the statistical models also resulted in high NU value, leading to under- or overestimation of annual mushroom productivity for certain ranges of the data. In GLMM, where the slope of the best linear fit between observed and predicted values was greater than 1, there was a clear underestimation. Moreover, GWR showed a slope between 0 and 1, leading to a production overestimation. On the contrary, machine learning models corrected this under- or overestimation by reducing the NU value and approximating the slope to 1 (Table S4).

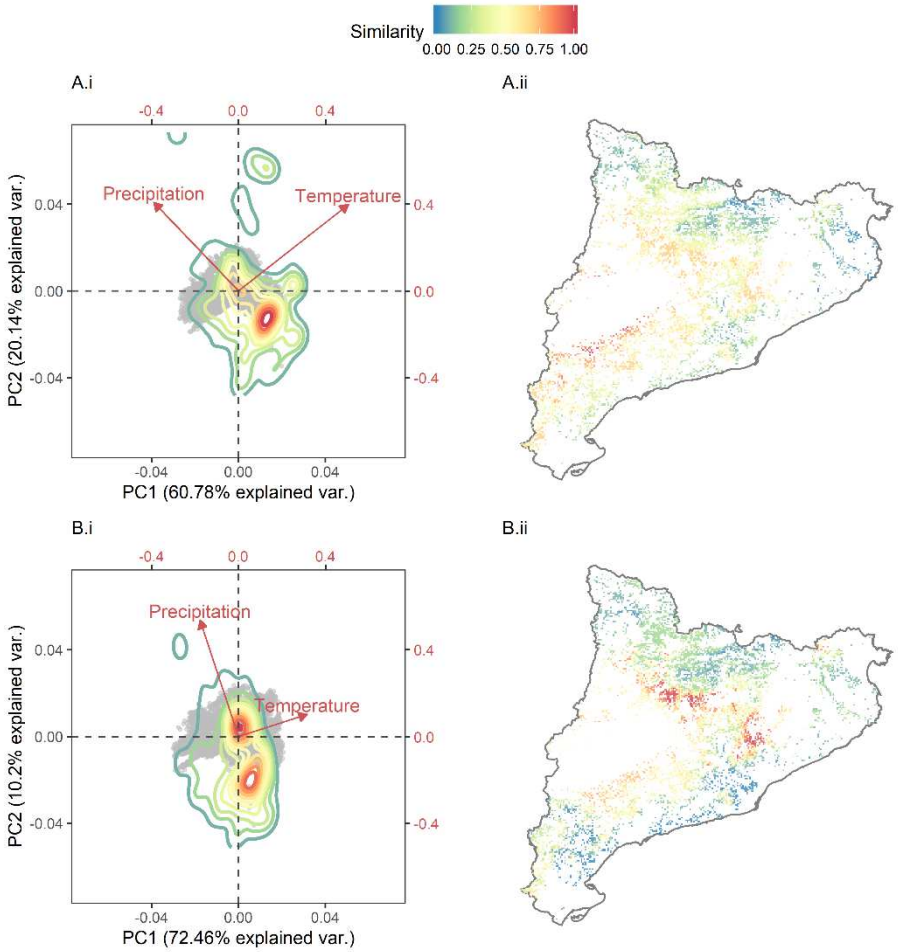


Figure 4. Similarity between the annual meteorological data of the sampled plots (used to train the models) and the average climatology of the Catalan pine forests (used to predict). Graphs show the climatology of the different pixels (grey) that form the map of pine forests in Catalonia from a two-dimensional representation based on a principal component analysis (PCA). The meteorological characteristics of the registered plots are shown from a density map of points within the bivariate space. Within this space, each pixel of pine forests matches a value of meteorological records density, being able to extrapolate this information to a map of the study region, obtaining a spatially explicit map of the similarity between these two datasets. (A.i) (A.ii) correspond to the similarities by creating a bivariate space resulting from PCA of 5 climatic variables (which can be used for the less dimensioned models), while (B.i) and (B.ii) are based on a PCA of 12 climate variables (which can be used in the models trained with 15 climatic variables).

3.3 Mapping and accuracy of predictions at the landscape level

The spatially explicit predictions at the landscape level from each model resulted in rather similar general patterns (Figure 5). Namely, they predicted higher productivity in the northern areas of the study region, characterized by higher altitudes, i.e., Pyrenees mountain range. Also, the different models reproduced similar patterns within these areas according to variations in local topography. In addition, RF, XGB and SVM models trained with 15 predictors (not in DL) yielded higher estimates of mushroom productivity in coastal areas compared to the same algorithms based on a subset of 5 predictors.

RF, XGB and SVM trained with 15 variables also resulted in less smoothed predictions of mushroom yield across the territory compared to estimates based on the subset of 5 predictors. Furthermore, SVM produced illogical predictions below 0 $kg \cdot ha^{-1} \cdot year^{-1}$ in a few spatially localized areas when 5 variables were used, and scattered throughout the territory when using 15 predictors. In contrast, DL resulted in very smoothed estimates across the territory, contrary to the maps obtained from all the other machine learning methods (Figure 5).

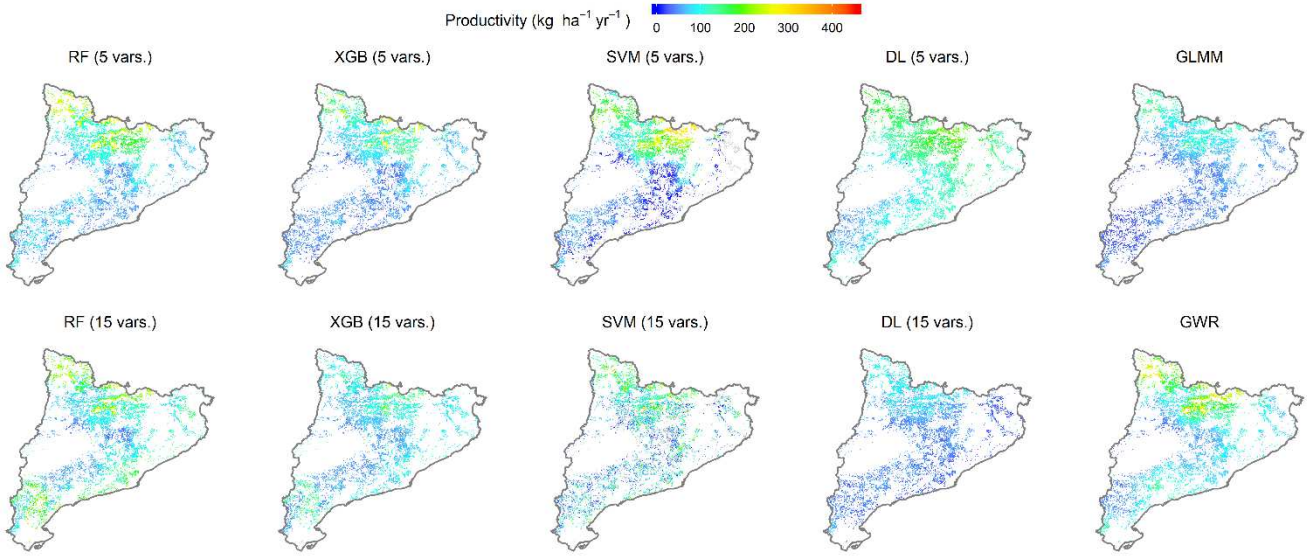


Figure 5. Landscape-level prediction of total annual mushroom productivity, using random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM)

and deep learning (DL) trained with 15 and 5 variables, respectively, and generalized linear mixed models (GLMM) and geographically weighted regression (GWR) fitted using 5 predictors.

In addition, the mushroom productivity predictions based on RF, XGB, SVM and GWR ranged between 0, in the less productive areas, and approximately 300 and 400 $kg \cdot ha^{-1} \cdot year^{-1}$. Slightly lower productivity was detected using GLMM and DL for the most productive sites, i.e. not exceeding 200 $kg \cdot ha^{-1} \cdot year^{-1}$ in any point of the study area.

ML models trained with 5 variables resulted in rather low relative SE in practically all the study area, whereas the models based on 15 predictors resulted in high relative SE in the coastal areas, i.e. those areas with the most different bioclimatic conditions compared to the training data. Finally, predictions using DL with 15 variables showed a high relative SE throughout the whole territory (Figure S7).

4 Discussion

To our knowledge, this is the first study addressing a systematic evaluation of the predictive performance of alternative statistical and machine learning models to predict mushroom productivity, and one of the few systematic comparisons between these different predictive approaches within the field of ecological research. This was conducted using one of the largest datasets (if not the largest one) for fungal productivity monitoring, based on consistent sampling methodology and taxonomic identification of mushrooms over more than 20 years on nearly a hundred permanent sampling plots, randomly distributed throughout the study region, which contributes to overcoming most of the practical problems related to the existence of available data for modelling fungal resources (Hao et al., 2020).

When dealing with complex ecological interactions between multiple potential explanatory variables, our results show that statistical models, specially GLMM, clearly seem to have lower predictive performance compared to artificial intelligence-based approaches, in line with previous research (e.g. Smoliński and Radtke (2016) and Schratz et al. (2019)). They were less accurate and produced large over- or underestimation of mushroom productivity, making them unreliable for such purposes compared to other alternatives. On the other hand, statistical models can be good candidates for detecting the most appropriate variables to be used in machine learning models and unravel environmental-ecological relationships between them (Shmueli, 2010; Schratz et al., 2019), since the inherent statistical assumptions that shape these models allow the relationships between data in a set of probability distributions to be correctly approximated. However, adding a spatial factor to the statistical models was able to correct for the strong underestimation of GLMM in high productivity values. By considering a spatial component, we were able to find stationary patterns in the predictors, denoting that climatic conditions do not affect equally across the study area.

As demonstrated here, choosing a subset of variables from statistically significant predictors in statistical models can help us to deal with some drawbacks. A problem with selecting a single subset of variables from a machine learning models is that, due to the algorithm itself, the significance is adjusted differently and could be inappropriate for some of them. For example, within decision tree algorithms, XGB determines the variable to be used in each node of the tree among the total of variables of a model, while RF does it within a subset of them, giving greater probability of being chosen to those less important variables (Hastie et al., 2001). On the other hand, and depending on the relationship between variables, the importance of a set of correlated variables can be distributed among the different predictors (giving lower importance to each one of them),

561 but the total importance that this set represents in the predictive performance is
562 remarkable (Toloşi and Lengauer, 2011). This can cause that when discarding the less
563 important variables, this set of predictors is omitted, causing a notorious drop in predictive
564 performance. Moreover, in a group of correlated variables where there is only one true
565 predictor (the one that implies real causality), machine learning algorithms could give
566 similar values of importance to the whole set of variables (Archer and Kime, 2007),
567 actually hiding the true predictor. Consequently, each machine learning algorithm give a
568 different importance to each variable. Therefore, to identify the variables that could best
569 explain the processes that occur in natural ecosystems and/or use the variable importance
570 to select a subset of predictors to train a more parsimonious model, the above
571 considerations should be taken into account.

572 Based on the prediction error of different models, it would seem logical to think that
573 models trained with 15 variables are better at achieving the goal of predicting mushroom
574 productivity (they have lower predictive error) compared to models based on five
575 predictors. However, they have some remarkable drawbacks. Using a larger amount of
576 variables one can account for more information about potential environmental drivers.
577 This might seem useful, because such information can help to improve predictions in
578 specific cases. However, using a larger number of variables also increases exponentially
579 the dimensions of the ecosystem space, commonly known as curse of dimensionality
580 Hughes (1968), which makes it more difficult to match the ranges of the modelling data
581 in extrapolation to a broader study area. In addition, not taking into account the data
582 spatial dependencies may increase the likelihood of getting out-of-bag misinterpretations
583 with increasing number of predictor variables (Meyer et al., 2018, 2019a). Conversely,
584 using fewer variables results in higher similarity (in terms of climatic conditions) between
585 the training and testing data. In particular, by creating a 2-D space based on PCA using

five climatic variables, the entire prediction zone was practically similar to the training data, while by using twelve climatic variables it decreases considerably (Figure 4). Compared to other alternative environmental blocking approaches (Valavi et al., 2018), our climatic blocking approach allowed us to identify the similarity between training and testing data as well as select the same amount of data in each blocking strategy. Assuming that spatial or climatic blocking are proper approaches to estimate the error in geographically (Roberts et al., 2017; Meyer et al., 2019a) or climatically different areas, and random blocking informs about the error in similar areas, we can determine which strategy will be best in each case. Knowing this, we can expect that random blocking may be more informative of the error of mushroom productivity estimates made by 5-variable models. On the other hand, in the spatial or climatic blockade it may be more correct to estimate the error of the estimates made by 15-variable models. In this way we can see that, in those areas where the error of the estimates of the 15-variable models is given by the spatial block and in the 5-variable models by the random block, a lower error is achieved with the second set of models. This is evident throughout the coastal zone of the study area and in the higher parts of the Pyrenees mountain range, where the error will be lower when using models with a smaller number of variables. Thus, it seems that parsimony can be useful model selection criterion not only for statistical methods, but also for machine learning algorithms (Coelho et al., 2018).

As noted, statistical models do not seem to be competitive compared to machine learning approaches due to poor predictive performance. Among the machine learning models, the DL approach had the highest prediction error and also resulted in biogeographic patterns that did not seem to agree with the expected climatic variations throughout the study area. In turn, SVM yielded illogical negative values of mushroom productivity in some areas. Therefore, the best candidate methods are the decision trees-based algorithms, i.e. RF and

XGB. Between both approaches, RF is preferable since it fits very well for this type of predictions, where the error mostly arises from the high variability of the studied ecological item (Hastie et al., 2001). This results in RF having the lowest error in all criteria analysed (Table S4), being better at estimating in sampling-like environments as in extrapolation. By estimating a value from the mean between the values given by a set of trees that are little, or not, correlated between them, it allows to reduce the variance. This study shows that, although machine learning algorithms allow, due to their characteristics, to fit models using a large number of variables, it may be advisable to make a selection of predictors to reduce the multidimensional space defined by the data to be predicted and thus become more similar to training data. This does not only allow us to address the selection of the best modelling approach to predict mushroom productivity; it also provides a methodology, due to the current paucity of data to build process-based models (Hao et al., 2020), that can be reasonably used in extrapolation out of the range of the modelling data. This is especially relevant in a context of global change, where climatic conditions are predicted to change over the years beyond the historical climatic ranges.

List of abbreviations

DL: Deep Learning

GLMM: Generalized Linear Mixed Models

GWR: Geographically Weighted Regression

MAE: Mean Absolute Error

PCA: Principal Component Analysis

PDP: Partial Dependence Plots

RF: Random Forest

RMSE: Root Mean Squared Error

SE: Standard Error

SVM: Support Vector Machine

XGB: Extreme Gradient Boosting

Author's contributions

JMdA, JAB and SdM contributed in the installation of the sampling plots and data collection. AM, SdM and JL analysed the data. AM wrote the manuscript. All authors participated in the review and editing of the manuscript. SdM supervised the work during the whole process.

Funding

This work was supported by the Secretariat for Universities and of the Ministry of Business and Knowledge of the Government of Catalonia and the European Social Fund. This work was partially supported by the Spanish Ministry of Science, Innovation and Universities, grant RTI2018-099315-A-I00. J.A.B. benefitted from a Serra-Húnter Fellowship provided by the Government of Catalonia.

Acknowledgements

Not applicable.

Availability of data and material

The datasets generated and/or analyzed during the current study are not publicly available due to legal constraints because they are owned by different institutions, but are available from the authors on reasonable request.

Ethics declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests.

References

Alday JG, Martínez de Aragón J, de-Miguel S, Bonet JA (2017) Mushroom biomass and diversity are driven by different spatio-temporal scales along Mediterranean elevation gradients. Scientific Reports, 7(1). doi:10.1038/srep45824

Allaire, JJ, Chollet, F (2019) keras: R Interface to 'Keras'. R package version 2.2.5.0. https://CRAN.R-project.org/package=keras

Archer KJ, Kimes RV (2008) Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52(4), 2249–2260. doi:10.1016/j.csda.2007.08.015

Bahn V, McGill BJ (2012) Testing the predictive performance of distribution models. Oikos, 122(3), 321–331. doi:10.1111/j.1600-0706.2012.00299.x

Barnard RL, Osborne CA, Firestone MK (2014) Changing precipitation pattern alters soil microbial community response to wet-up under a Mediterranean-type climate. The ISME Journal, 9(4), 946–957. doi:10.1038/ismej.2014.192

Bastin J-F, Finegold Y, Garcia C, Mollicone D, Rezende M, Routh D, Constantin MZ, Crowther TW (2019) The global tree restoration potential. Science, 365(6448), 76–79. doi:10.1126/science.aax0848

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

Bivand R, Yu D (2017) spgwr: Geographically Weighted Regression. R package version 0.6-32. https://CRAN.R-project.org/package=spgwr

697 Boa E. (2004) Wild edible fungi: a global overview of their use and importance to people.
698 Non-Wood Forest Products, No. 17, FAO. Forestry Department, Rome, Italy, 148p.
699 ISBN: 92-5-105157-7

700 Bonet JA, Fischer CR, Colinas C (2004) The relationship between forest age and aspect
701 on the production of sporocarps of ectomycorrhizal fungi in Pinus sylvestris forests
702 of the central Pyrenees. Forest Ecology and Management, 203(1-3), 157–175.
703 doi:10.1016/j.foreco.2004.07.063

704 Bonet JA, Palahí M, Colinas C, Pukkala T, Fischer CR, Miina J, Martínez de Aragón, J
705 (2010) Modelling the production and species richness of wild mushrooms in pine
706 forests of the Central Pyrenees in northeastern Spain. Canadian Journal of Forest
707 Research, 40(2), 347–356. doi:10.1139/x09-198

708 Bonete IP, Arce JE, Figueiredo Filho A, Retslaff FA de S, Lanssanova LR (2020)
709 Artificial neural networks and mixed-effects modeling to describe the stem profile of
710 Pinus taeda l. floresta, 50(1), 1123. doi:10.5380/rf.v50i1.61764

711 Breiman L (2001) Random Forest. Machine Learning, 45(1), 5–32.
712 doi:10.1023/a:1010933404324

713 Büntgen U, Kauserud H, Egli S (2012) Linking climate variability to mushroom
714 productivity and phenology. Frontiers in Ecology and the Environment, 10(1), 14–
715 19. doi:10.1890/110064

716 Büntgen U, Peter M, Kauserud H, Egli S (2013) Unraveling environmental drivers of a
717 recent increase in Swiss fungi fruiting. Global Change Biology, 19(9), 2785–2794.
718 doi:10.1111/gcb.12263

719 Chen T, Guestrin C (2016) KDD '16: Proceedings of the 22nd ACM SIGKDD
720 International Conference on Knowledge Discovery and Data Mining, August 2016,
721 Pages 785–794 doi:10.1145/2939672.2939785

722 Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, … Li Y (2019) xgboost:
723 Extreme Gradient Boosting. R package version 0.90.0.2. https://CRAN.R-
724 project.org/package=xgboost

725 Christin S, Hervet É, Lecomte N (2019) Applications for deep learning in ecology.
726 Methods Ecol Evol, 10:1632–1644. doi:10.1111/2041- 210X.13256

727 Coelho MTP, Diniz- Filho JA, Rangel TF (2018) A parsimonious view of the parsimony
728 principle in ecology and evolution. Ecography. doi:10.1111/ecog.04228

729 Collado E, Camarero JJ, Martínez de Aragón J, Pemán J, Bonet JA, de-Miguel S (2018)
730 Linking fungal dynamics, tree growth and forest management in a Mediterranean
731 pine ecosystem. Forest Ecology and Management, 422, 223–232.
732 doi:10.1016/j.foreco.2018.04.025

733 Collado E, Bonet JA, Camarero JJ, Egli S, Peter M, Salo K, … de-Miguel S (2019)
734 Mushroom productivity trends in relation to tree growth and climate across different
735 European forest biomes. Science of The Total Environment.
736 doi:10.1016/j.scitotenv.2019.06.471

737 Commission of the European Community. (1991) CORINE biotopes manual – Habitats
738 of the European Community. Luxembourg: DG Environment, Nuclear Safety and
739 Civil Protection.

740 Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning, 20(3), 273–297.
741 doi:10.1007/bf00994018

742 Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007)
743 Random forests for classification in ecology. Ecology, 88: 2783-2792.
744 doi:10.1890/07-0539.1

745 De Cáceres M, Martin-StPaul N, Turco M, Cabon A, Granda V (2018) Estimating daily
746 meteorological data and downscaling climate models over landscapes.
747 Environmental Modelling & Software, 108, 186–196.
748 doi:10.1016/j.envsoft.2018.08.003

749 de-Miguel S, Bonet JA, Pukkala T, Martínez de Aragón J (2014) Impact of forest
750 management intensity on landscape-level mushroom productivity: A regional model-
751 based scenario analysis. Forest Ecology and Management, 330, 218–227.
752 doi:10.1016/j.foreco.2014.07.014

753 Diamantopoulou MJ, Özçelik R, Crecente-Campo F, Eler Ü (2015) Estimation of Weibull
754 function parameters for modelling tree diameter distribution using least squares and
755 artificial neural networks methods. Biosyst. Eng. 133:33–45.
756 doi:10.1016/j.biosystemseng.2015.02.013

757 Dixon PA, Milicich MJ, Sugihara G (1999) Episodic Fluctuations in Larval Supply.
758 Science, 283(5407), 1528–1530. doi:10.1126/science.283.5407.1528

759 Ehrlén J, Morris WF (2015) Predicting changes in the distribution and abundance of
760 species under environmental change. Ecology Letters, 18(3), 303–314.
761 doi:10.1111/ele.12410

762 Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view
763 of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics,
764 28(2), 337–407. doi:10.1214/aos/1016218223

765 Gange AC, Gange EG, Sparks TH, Boddy L (2007) Rapid and Recent Changes in Fungal
766 Fruiting Patterns. Science, 316(5821), 71–71. doi:10.1126/science.1137489

767 Gasch CK, Hengl T, Gräler B, Meyer H, Magney TS, Brown DJ (2015) Spatio-temporal
768 interpolation of soil water, temperature, and electrical conductivity in 3D + T: The
769 Cook Agronomy Farm data set. Spatial Statistics, 14, 70–90.
770 doi:10.1016/j.spasta.2015.04.001

771 Gauch HG, Hwang JTG, Fick GW (2003) Model Evaluation by Comparison of Model-
772 Based Predictions and Measured Values. Agronomy Journal, 95(6), 1442.
773 doi:10.2134/agronj2003.1442

774 Georganos S, Abdi AM, Tenenbaum DE, Kalogirou S (2017) Examining the NDVI-
775 rainfall relationship in the semi-arid Sahel using geographically weighted regression.
776 Journal of Arid Environments, 146, 64–74. doi:10.1016/j.jaridenv.2017.06.004

777 Glassman SI, Wang IJ, Bruns TD (2017) Environmental filtering by pH and soil nutrients
778 drives community assembly in fungi at fine spatial scales. Mol. Ecol., 26, pp. 6960-
779 6973. doi:10.1111/mec.14414

780 Gobeyn S, Mouton AM, Cord AF, Kaim A, Volk M, Goethals PLM (2019) Evolutionary
781 algorithms for species distribution modelling: A review in the context of machine
782 learning. Ecological Modelling, 392, 179–195.
783 doi:10.1016/j.ecolmodel.2018.11.013

784 Görgens EB, Montaghi A, Rodriguez LCE (2015) A performance comparison of machine
785 learning methods to estimate the fast-growing forest plantation yield based on laser
786 scanning metrics. Computers and Electronics in Agriculture, 116, 221–227.
787 doi:10.1016/j.compag.2015.07.004

788 Hamilton Jr D A, Brickell JE (1983) Modeling methods for a two-state system with
789 continuous responses. Canadian Journal of Forest Research, 13(6), 1117–1121.
790 doi:10.1139/x83-149

791 Hannemann H, Willis KJ, Macias-Fauria M (2015) The devil is in the detail: unstable
792 response functions in species distribution models challenge bulk ensemble
793 modelling. Global Ecology and Biogeography, 25(1), 26–35. doi:10.1111/geb.12381

794 Hao T, Guillera-Arroita G, May TW, Lahoz-Monfort JJ, Elith J (2020) Using Species
795 Distribution Models for fungi. Fungal Biology Reviews.
796 doi:10.1016/j.fbr.2020.01.002

797 Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data
798 Mining, Inference, and Prediction. New York: Springer-Verlag. ISBN 0-387-95284-
799 5

800 Hill L, Hector A, Hemery G, Smart S, Tanadini M, Brown N (2017) Abundance
801 distributions for tree species in Great Britain: A two-stage approach to modeling
802 abundance using species distribution modeling and random forest. Ecology and
803 Evolution, 7: 1043–1056. doi:10.1002/ece3.2661

804 Hughes G (1968) On the mean accuracy of statistical pattern recognizers. IEEE
805 Transactions on Information Theory, 14(1), 55–63. doi:10.1109/tit.1968.1054102

806 James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning.
807 Springer Texts in Statistics. doi:10.1007/978-1-4614-7138-7

808 Juel A, Groom GB, Svenning J-C, Ejrnæs R (2015) Spatial application of Random Forest
809 models for fine-scale coastal vegetation classification using object based analysis of

aerial orthophoto and DEM data. International Journal of Applied Earth Observation and Geoinformation, 42, 106–114. doi:10.1016/j.jag.2015.05.008

Karavani A, De Cáceres M, Martínez de Aragón J, Bonet JA, de-Miguel S (2018) Effect of climatic and soil moisture conditions on mushroom productivity and related ecosystem services in Mediterranean pine stands facing climate change. Agricultural and Forest Meteorology, 248, 432–440. doi:10.1016/j.agrformet.2017.10.024

Kauserud H, Stige LC, Vik JO, Okland RH, Hoiland K, Stenseth NC (2008) Mushroom fruiting and climate change. Proceedings of the National Academy of Sciences, 105(10), 3811–3814. doi:10.1073/pnas.0709037105

Kauserud H, Heegaard E, Semenov MA, Boddy L, Halvorsen R, Stige LC, … Stenseth, NC (2009) Climate change and spring-fruiting fungi. Proceedings of the Royal Society B: Biological Sciences, 277(1685), 1169–1177. doi:10.1098/rspb.2009.1537

Le Rest K, Pinaud D, Monestiez P, Chadoeuf J, Bretagnolle V (2014) Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecology and Biogeography, 23(7), 811–820. doi:10.1111/geb.12161

Liang J, Crowther TW, Picard N, Wiser S, Zhou M, Alberti G, … Pretzsch H (2016) Positive biodiversity-productivity relationship predominant in global forests. Science, 354(6309), aaf8957–aaf8957. doi:10.1126/science.aaf8957

Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2(3), 18--22

Martínez de Aragón J, Bonet JA, Fischer CR, Colinas C (2007) Productivity of ectomycorrhizal and selected edible saprotrophic fungi in pine forests of the pre-Pyrenees mountains, Spain: Predictive equations for forest management of

833 mycological resources. Forest Ecology and Management, 252(1-3), 239–256.
834 doi:10.1016/j.foreco.2007.06.040

835 Mayfield H, Smith C, Gallagher M, Hockings M (2020) Considerations for selecting a
836 machine learning technique for predicting deforestation. Environmental Modelling
837 Software, 104741. doi:10.1016/j.envsoft.2020.104741

838 Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T (2018) Improving performance of
839 spatio-temporal machine learning models using forward feature selection and target-
840 oriented validation. Environmental Modelling Software, 101, 1–9.
841 doi:10.1016/j.envsoft.2017.12.001

842 Meyer H, Reudenbach C, Wöllauer S, Nauss T. (2019a) Importance of spatial predictor
843 variable selection in machine learning applications – Moving from data reproduction
844 to spatial prediction. Ecological Modelling, 411, 108815.
845 doi:10.1016/j.ecolmodel.2019.108815

846 Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2019b) e1071: Misc
847 Functions of the Department of Statistics, Probability Theory Group (Formerly:
848 E1071), TU Wien. R package version 1.7-2. https://CRAN.R-
849 project.org/package=e1071

850 Micheletti N, Foresti L, Robert S, Leuenberger M, Pedrazzini A, Jaboyedoff M, Kanevski
851 M (2013) Machine Learning Feature Selection Methods for Landslide Susceptibility
852 Mapping. Mathematical Geosciences, 46(1), 33–57. doi:10.1007/s11004-013-9511-
853 0

854 Mohan JE, Cowden CC, Baas P, Dawadi A, Frankson PT, Helmick K, … Witt CA (2014)
855 Mycorrhizal fungi mediation of terrestrial ecosystem responses to global change:
856 mini-review. Fungal Ecology, 10, 3–19. doi:10.1016/j.funeco.2014.01.005

857 Olden JD, Lawler JJ, Poff NL (2008) Machine Learning Methods Without Tears: A
858 Primer for Ecologists. The Quarterly Review of Biology, 83(2), 171–193.
859 doi:10.1086/587826

860 Özçelik R, Diamantopoulou MJ, Crecente-Campo F, Eler U (2013) Estimating Crimean
861 juniper tree height using nonlinear regression and artificial neural network models.
862 Forest Ecology and Management. 306, 52–60. doi:10.1016/j.foreco.2013.06.009

863 Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS. First ed. Springer,
864 New York. doi:10.1007/b98882

865 Pohjankukka J, Pahikkala T, Nevalainen P, Heikkonen J (2017) Estimating the prediction
866 performance of spatial models via spatial k-fold cross validation. International
867 Journal of Geographical Information Science, 31(10), 2001–2019.
868 doi:10.1080/13658816.2017.1346255

869 Prasad A, Iverson L, Liaw A (2006) Newer Classification and Regression Tree
870 Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems 9
871 (2): 181 199. doi:10.1007/s10021-005-0054-1

872 Prasad AM (2018) Machine Learning for Macroscale Ecological Niche Modeling - a
873 Multi-Model, Multi-Response Ensemble Technique for Tree Species Management
874 Under Climate Change. Machine Learning for Ecology and Sustainable Natural
875 Resource Management, 123–139. doi:10.1007/978-3-319-96978-7_6

876 Propastin PA (2009) Spatial non-stationarity and scale-dependency of prediction
877 accuracy in the remote estimation of LAI over a tropical rainforest in Sulawesi,
878 Indonesia. Remote Sensing of Environment, 113(10), 2234–2242.
879 doi:10.1016/j.rse.2009.06.007

880 Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Dormann CF (2017)
881 Cross-validation strategies for data with temporal, spatial, hierarchical, or
882 phylogenetic structure. Ecography, 40(8), 913–929. doi:10.1111/ecog.02881

883 Sánchez-González M, de-Miguel S, Martin-Pinto P, Martínez-Peña F, Pasalodos-Tato M,
884 Oria-de-Rueda JA, … Bonet JA (2019) Yield models for predicting aboveground
885 ectomycorrhizal fungal productivity in Pinus sylvestris and Pinus pinaster stands of
886 northern Spain. Forest Ecosystems, 6(1) doi:10.1186/s40663-019-0211-1

887 Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A (2019) Hyperparameter tuning
888 and performance assessment of statistical and machine-learning algorithms using
889 spatial data. Ecological Modelling, 406, 109–120.
890 doi:10.1016/j.ecolmodel.2019.06.002

891 Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier
892 performance in R. Bioinformatics, 21(20), pp. 7881. http://rocr.bioinf.mpi-sb.mpg.de

893 Shmueli G (2010) To explain or to predict? Statistical Science 25(3), 289-310
894 doi:10.1214/10-STS330

895 Smoliński S, Radtke K (2016) Spatial prediction of demersal fish diversity in the Baltic
896 Sea: comparison of machine learning and regression-based techniques. ICES Journal
897 of Marine Science: Journal Du Conseil, fsw136. doi:10.1093/icesjms/fsw136

898 Snowdon P (1991) A ratio estimator for bias correction in logarithmic regressions.
899 Canadian Journal of Forest Research, 21(5), 720–724. doi:10.1139/x91-101

900 Stojanova D, Panov P, Gjorgjioski V, Kobler A, Džeroski S (2010) Estimating vegetation
901 height and canopy cover from remotely sensed data with machine learning.
902 Ecological Informatics, 5(4), 256–266. doi:10.1016/j.ecoinf.2010.03.004

903 Stokland JN, Siitonen J, Jonsson BG (2012) Biodiversity in Dead Wood, Biodiversity in
904 Dead Wood. Cambridge University Press, Cambridge.
905 doi:10.1017/CBO9781139025843

906 Taye ZM, Martínez-Peña F, Bonet JA, Martínez de Aragón J, de-Miguel S (2016)
907 Meteorological conditions and site characteristics driving edible mushroom
908 production in Pinus pinaster forests of Central Spain. Fungal Ecology, 23, 30–41.
909 doi:10.1016/j.funeco.2016.05.008

910 Thessen A (2016) Adoption of Machine Learning Techniques in Ecology and Earth
911 Science. One Ecosystem, 1: e8621. doi:10.3897/oneeco.1.e8621

912 Thornton PE, Running SW, White MA (1997) Generating surfaces of daily
913 meteorological variables over large regions of complex terrain. Journal of
914 Hydrology, 190(3-4), 214–251. doi:10.1016/s0022-1694(96)03128-9

915 Thornton PE, Running SW (1999) An improved algorithm for estimating incident daily
916 solar radiation from measurements of temperature, humidity, and precipitation.
917 Agricultural and Forest Meteorology, 93(4), 211–228. doi:10.1016/s0168-
918 1923(98)00126-9

919 Thuiller W (2003) BIOMOD - optimizing predictions of species distributions and
920 projecting potential future shifts under global change. Global Change Biology, 9
921 (10), 1353– 1362. doi:10.1046/j.1365-2486.2003.00666.x

922 Toloşi L, Lengauer T (2011) Classification with correlated features: unreliability of
923 feature ranking and solutions. Bioinformatics, 27(14), 1986–1994.
924 doi:10.1093/bioinformatics/btr300

925 Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2018) blockCV: an R package
926 for generating spatially or environmentally separated folds for k-fold cross-validation
927 of species distribution models. Methods in Ecology and Evolution.
928 doi:10.1111/2041-210x.13107

929 Wheeler DC, Páez A (2009) Geographically Weighted Regression. In: Fischer M, Getis
930 A. (eds) Handbook of Applied Spatial Analysis. Springer, Berlin, Heidelberg.
931 doi:10.1007/978-3-642-03647-7_22

932 Wood SN, Thomas MB (1999) Super-sensitivity to structure in biological models. Proc
933 R Soc Lond B Biol Sci 266(1419):565–570. doi:10.1098/rspb.1999.0673

934 Ye H, Beamish RJ, Glaser SM, Grant SC, Hsieh C, Richards LJ, Schnute JT, Sugihara G
935 (2015) Equation-free mechanistic ecosystem forecasting using empirical dynamic
936 modeling. Proceedings of the National Academy of Sciences, 112:E1569–1576
937 doi:10.1073/pnas.1417063112

938

939

940

941

942

943

944

945

946

Supplementary Material

3.1 Fitted models

3.1.1 Statistical models

Here we show the fitted GLMM and GWR models (coefficients are in Table S1 to S3) It should be noted that the coefficients of GLMM and GWR are particular and different from the general form of the conditioned mushroom production (S1) and the logit transformation of the logistic regression (S2)

$$\ln(prod) = \beta_0 + \beta_1 \ln\left(P_{aug+sep+oct}\right) + \beta_2 \ln(Tmax_{set}) \quad \text{(S1)}$$

$$g = \beta_3 + \beta_4 \ln\left(P_{aug+sep+oct}\right) + \beta_5 Tmax_{oct} \quad \text{(S2)}$$

Table S1. Fixed GLMM coefficients. β1, β2, β3 significance were calculated from t-value, while β3, β4, β5 from p-value.

Coefficients	Estimate	Std. Error	Significance
β_0	2.858	2.203	1.298
β_1	1.048	0.156	6.680
β_2	-1.535	0.628	-2.442
β_3	6.437	4.999	0.198
β_4	2.149	0.598	0.000
β_5	-2.591	0.961	0.007

962 Table S2. Random GLMM coefficients.

Coefficient	Minimum	1st quantile	Median	3rd quantile	Maximum
β_0	-1.748	-0.408	0.096	0.425	1.323
β_3	-2.043	-0.395	0.359	0.821	1.363

963

964 Table S3. GWR models coefficients.

Coefficient	Minimum	1st quantile	Median	3rd quantile	Maximum
β_0	-4.207	-2.932	-0.612	0.298	0.980
β_1	1.439	1.580	1.738	2.231	2.636
β_2	-2.233	-1.724	-1.482	-1.398	-0.786
β_3	-7.908	-5.375	1.853	7.018	29.509
β_4	0.736	1.177	1.830	2.610	3.517
β_5	-7.096	-1.775	-1.453	-1.261	-0.654

Table S4. Model error in terms of root mean squared error (RMSE) and mean squared error (MSE) and his decomposition in squared bias (SB), nonunity slope(NU) and lack of correlation (LC) for random forest models (RF), extreme gradient (XGB), support vector machine (SVM), deep learning (DL), generalized linear mixed models (GLMM) and geographically weighted regression (GWR) models trained with 5 (.5) and 15 (.15 variables.

Blocking	% testing data	Parameter	RF.5	RF.15	XGB.5	XGB.15	SVM.5	SVM.15	DL.5	DL.15	GLMM	GWR
Random Blocking	5	SB	3.79	17.74	201.47	210.37	484.1	223.31	5.55	22.26	3169	22.89
	5	NU	99.25	81.29	29.6	105.46	5.14	43.37	0.25	3.87	2078.3	1544.68
	5	LC	12306.74	11138.82	12133.91	11405.9	13056.35	12560.4	14896.03	14741.47	15819.7	14416.35
	5	MSE	12409.78	11237.85	12364.98	11721.73	13545.59	12827.08	14901.83	14767.61	21067.01	15983.93
	5	RMSE	113.59	105.24	118.59	114.65	122.94	116.26	132.04	131.35	153.19	130.1
	10		113.52	108.54	118.02	114.32	119.78	116.8	127.88	127.01	147.92	127.35
	20		113.29	108.22	117.89	114.29	117.15	114.78	125.64	124.72	144.19	127.84
Spatial Blocking	5		139.13	133.9	149.22	136.4	137.64	138.06	140.59	136.49	166.95	153.48
	10		123.75	122.6	128.98	126.96	133.67	135.66	125.32	125.12	138.22	134.84
	20		124.18	124.84	128.66	126.52	135.65	137.63	126.39	126.04	135.1	133.93
Climate Blocking	5		111.46	110.24	114.96	111.97	120.08	116.51	115.7	115.6	128.12	128.52
	10		128.3	126.13	137.36	129.71	144.76	133.88	130.8	132.37	143.22	142.02
	20		123.17	121.73	126.92	126.13	133.99	129.01	126.74	127.46	134.44	134.91
Substitution	5		58.83	51.58	88.41	81.87	105.97	84.12	124.38	123.5	143.73	128.01

Supplementary figures

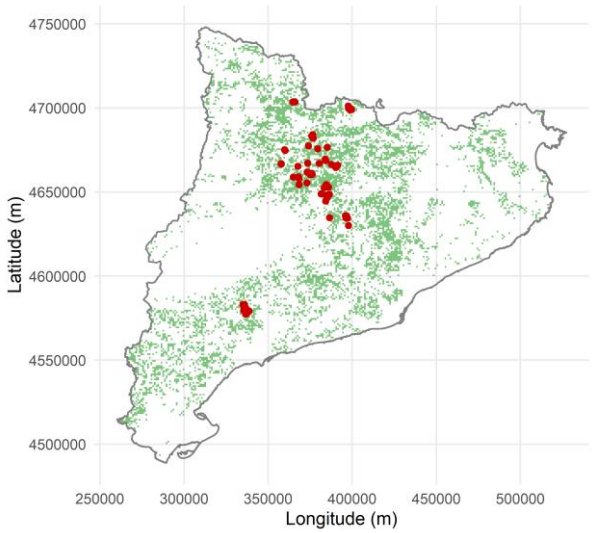
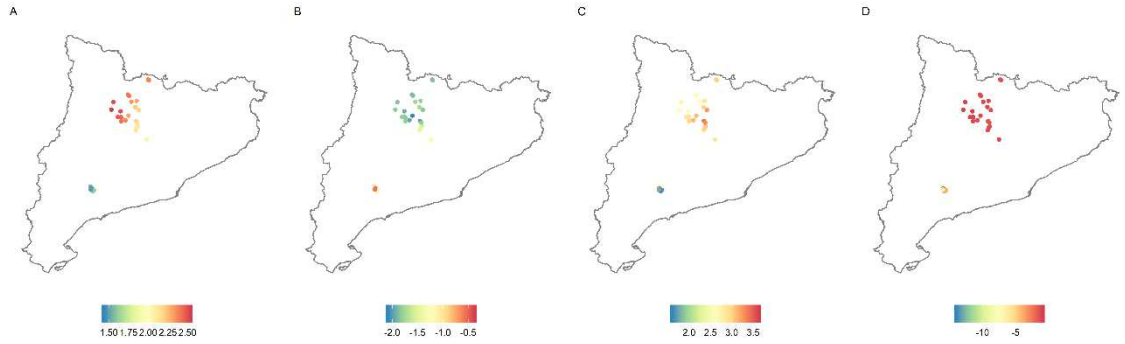


Figure S1. Study area, distribution of mushroom productivity monitoring plots (red points) and pine forest ecosystems represented by the sample plots (green area) Coordinates system: WGS 84 / UTM zone 31N.



974 Figure S2. GWR coefficient estimates according to geographical location. Coefficient of
975 precipitation amount from August to October (A) and maximum temperature in August
976 (B) in conditioned production model. Coefficient of precipitation amount from August to
977 October (A) and maximum temperature in October (D) in occurrence model (C).

978

979

980

981

982

983

984

985

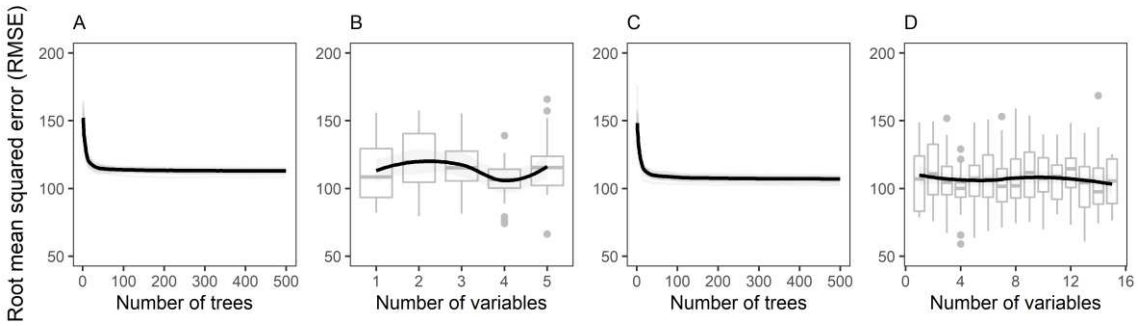
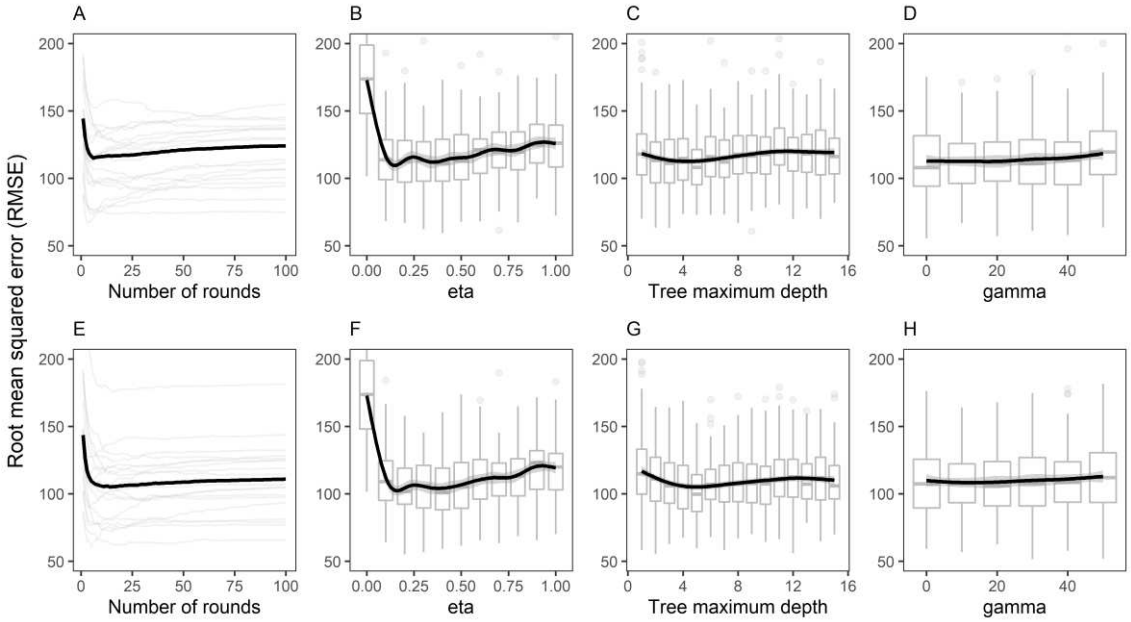


Figure S3. Hyperparameters selection for Random forest algorithm. (A) and (B) corresponds to random forest models trained with 5 variables while (C) and (D) to those trained with 15. For the trained models with 5 and 15 variables we chose 200 trees per each model and 1 and 5 variables to be chosen randomly in each division, respectively.



997

998 Figure S4. Hyperparameters selection for Extreme gradient boosting algorithm. (A) to
999 (D) corresponds to extreme gradient boosting models trained with 5 variables white (E)
1000 to (H) to those trained with 15. We chose 20 number of rounds and an eta of 0.1 for each
1001 model while choosing a depth of 5 and no regularization (gamma=0).

1002

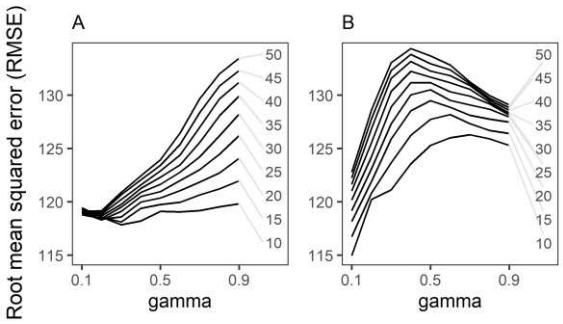
1003

1004

1005

1006

1007



1008

1009 Figure S5. Hyperparameters selection for Support vector machine algorithm. Different
1010 lines correspond to the value of the hyperparameter cost. (A) refers to support vector
1011 machine models trained with 5 variables while (B) to those trained with 15. We chose a
1012 gamma value of 0.3 and a cost of 10 for the 5-variable model while 0.1 and 10 respectively
1013 for the 15-variable model.

1014

1015

1016

1017

1018

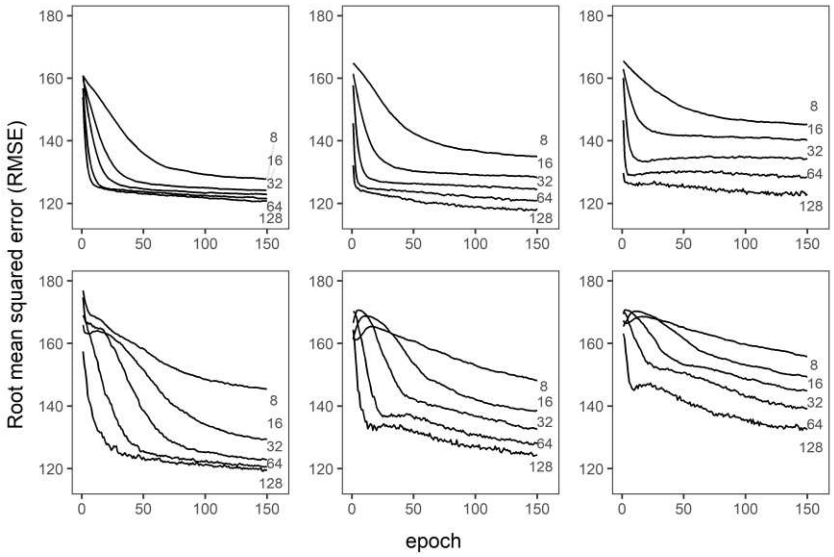
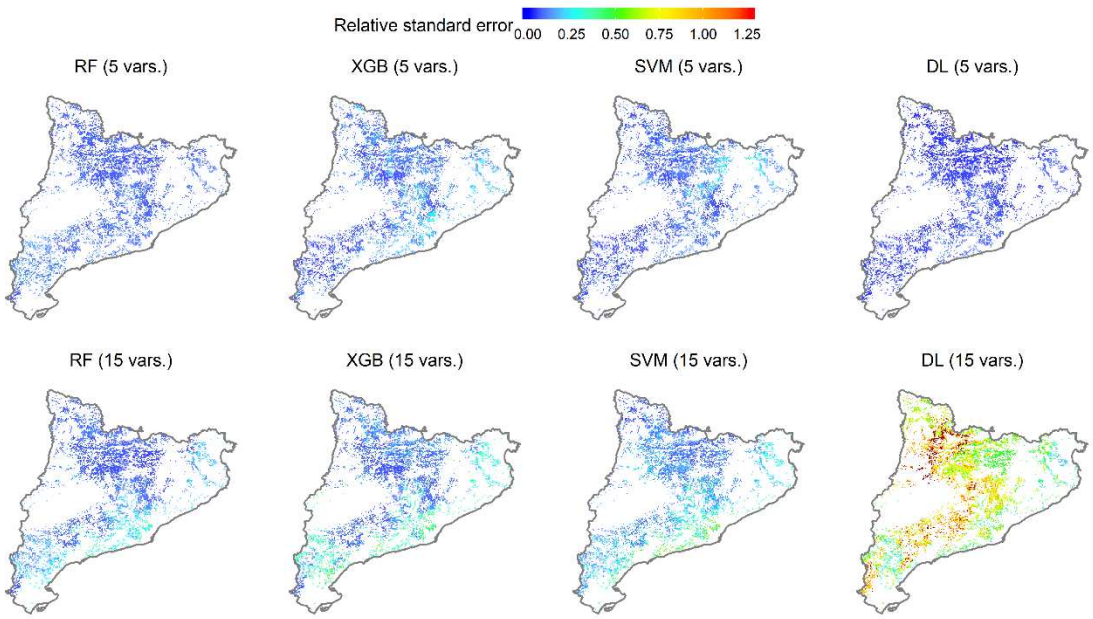


Figure S6. Hyperparameters selection for Deep learning algorithm. (A) to (C) refers to deep learning models trained with 5 variables while (D) to (F) to those trained with 15. (A) and (D) show ferments per 1 each with an activation function "relu" while (B) and (E) 2 and (C) and (F) 3. A final layer with a single output value and a "relu" activation function is added to all of them. In between each layer a dropout layer has been added which randomly selects 50% of the data for training. The different lines of each graph show the dimensionality of the output space.



1027 Figure S7. Spatially explicit standard error map of the 50 model estimations

Figures

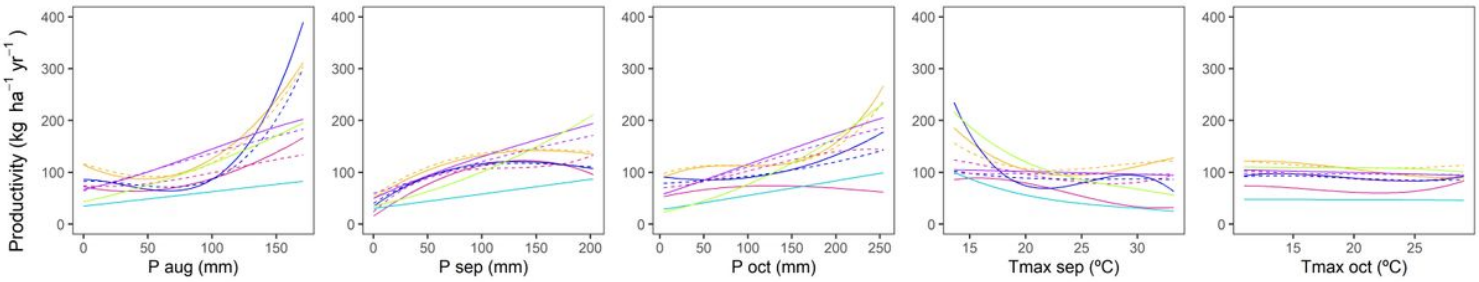


Figure 1

Relationship between annual mushroom productivity and August, September and October precipitation and maximum temperatures in September and October (these variables are the variables used in the statistical models and the five variables machine learning models). Yellow (random forest), blue (extreme gradient boosting), violet (support vector machine), purple (deep learning), light blue (generalized linear mixed models) and light green (geographically weighted regression) colours refer to the different modelling techniques. Continuous line includes the models that use the five variables and the dashed line the machine learning models trained with 15.

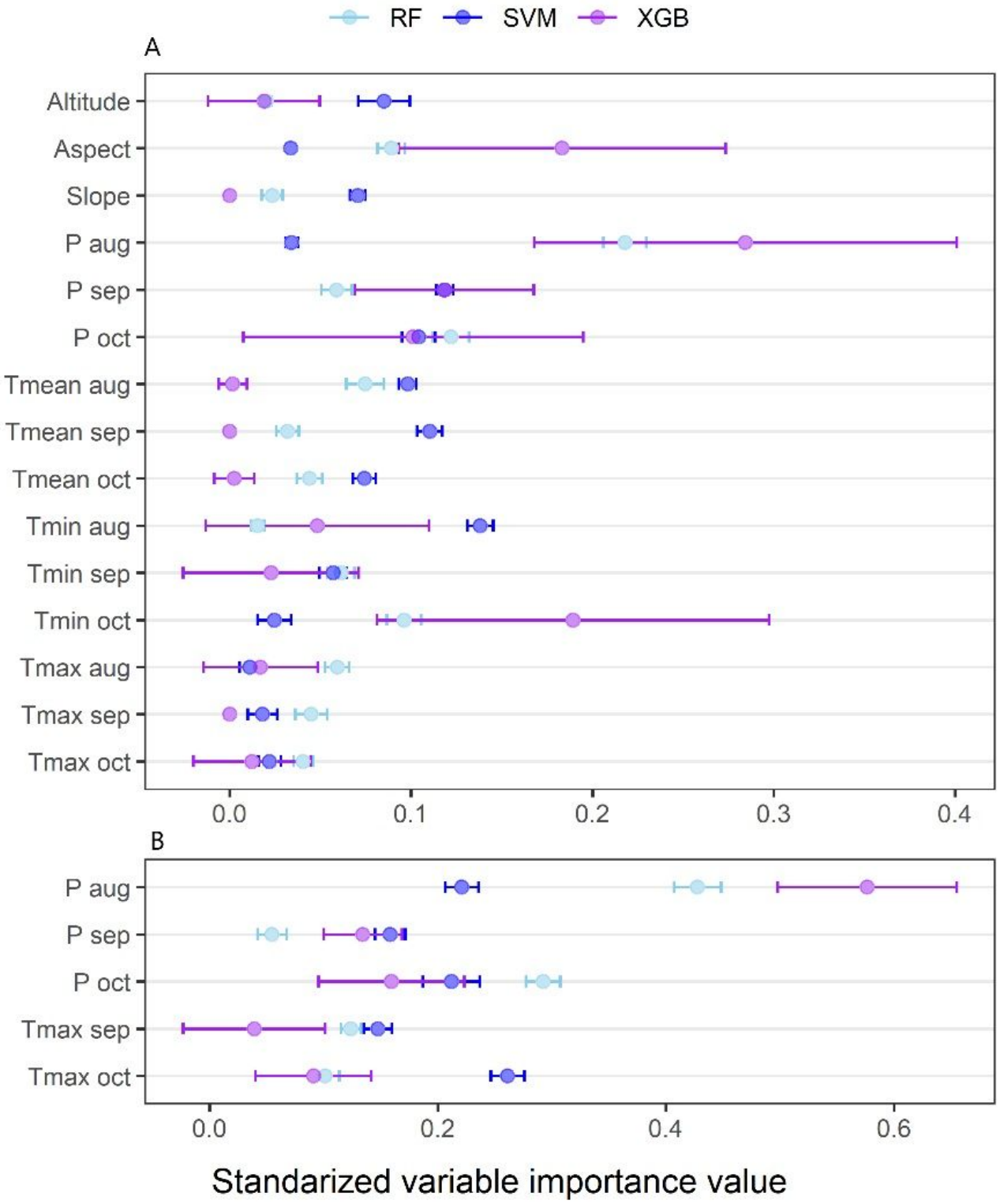


Figure 2

Standardized variable importance value used to train random forest (RF), extreme gradient boosting (XGB) and support vector machine (SVM) models with 15 (A) and 5 (B) variables. Variable importance values represent the contribution of each variable to predict the annual mushroom productivity.

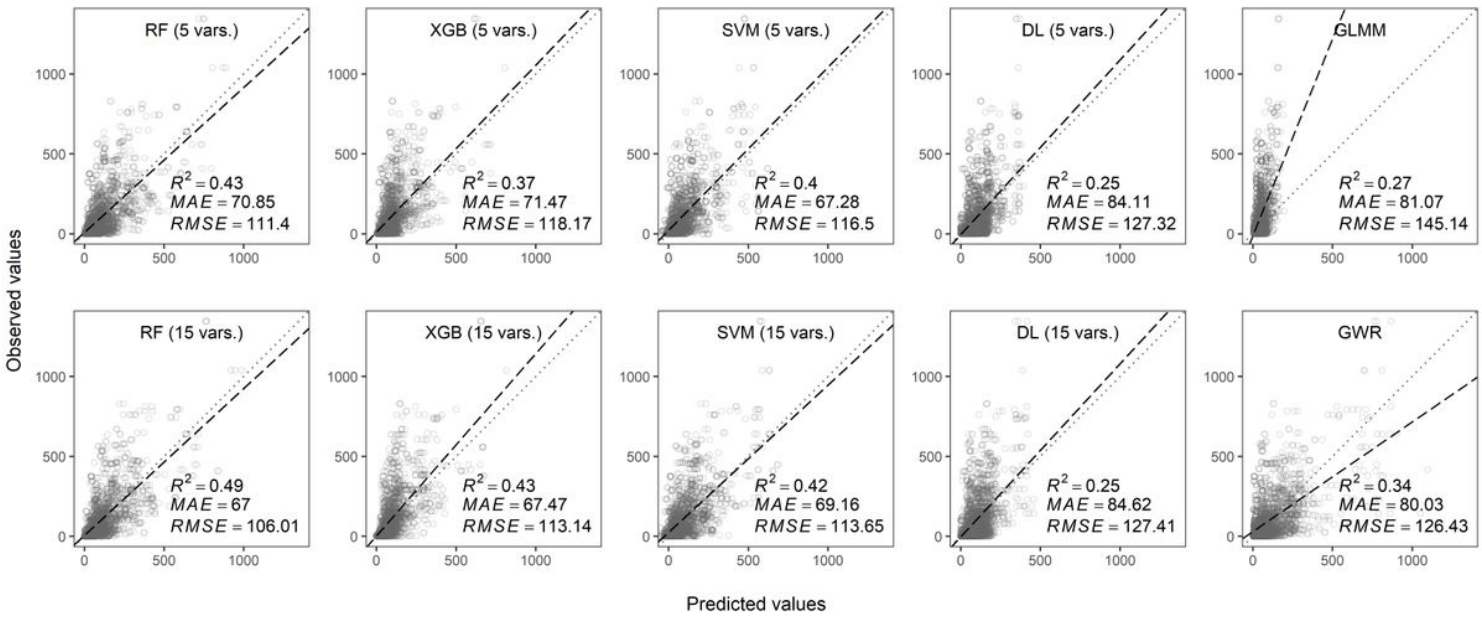


Figure 3

Cross-validation of estimated total mushroom production (kg ha-1 year-1) using random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and deep learning (DL) trained with 15 and 5 variables and generalized linear mixed models (GLMM) and geographically weighted regression (GWR). Pointed line shows the equality line and the dashed one the best fit between predicted and observed values. It shows mean absolute error (MAE), root-mean-squared error (RMSE) and coefficient of determination (R2) of the relation for each technique.

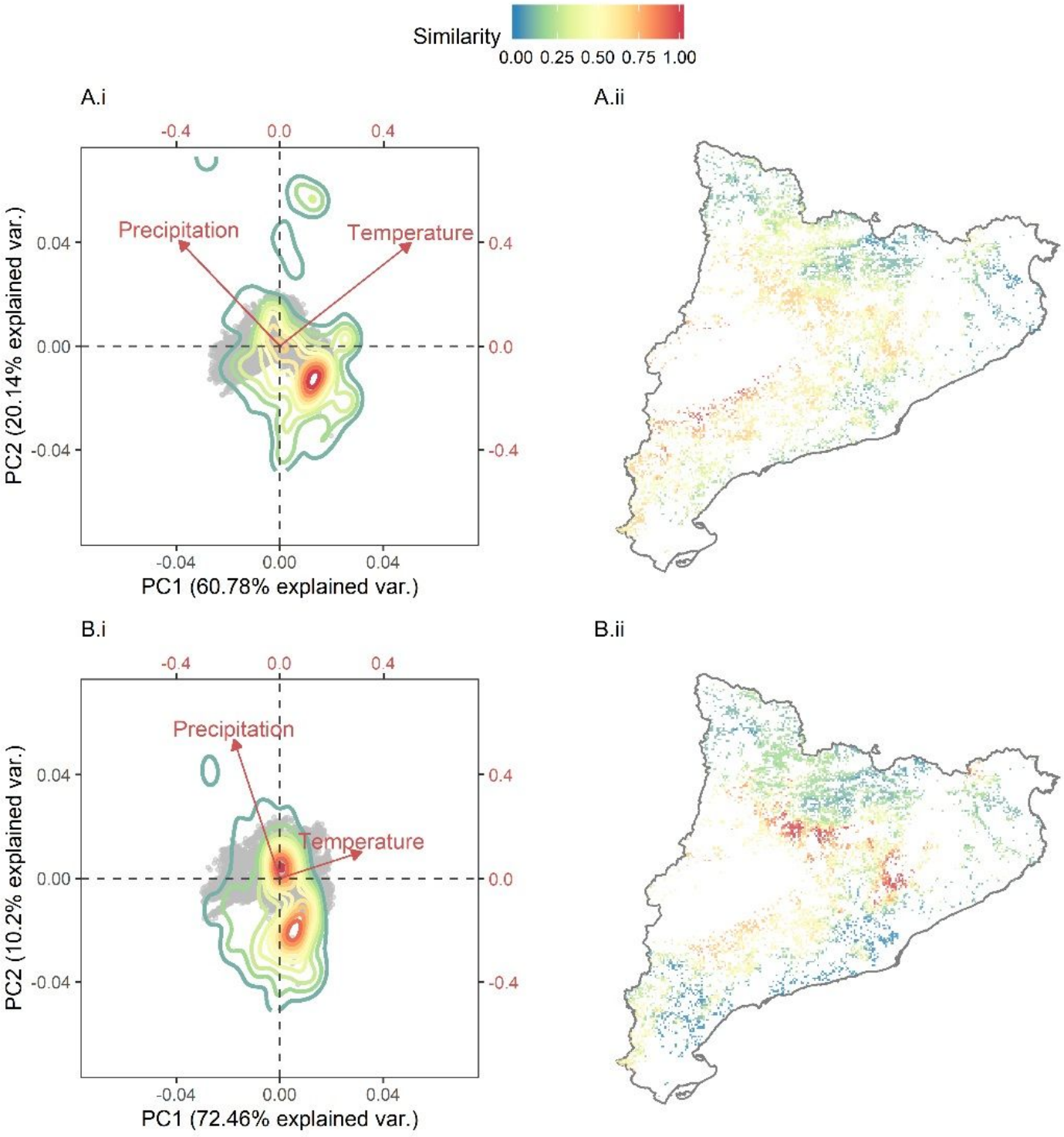


Figure 4

Similarity between the annual meteorological data of the sampled plots (used to train the models) and the average climatology of the Catalan pine forests (used to predict).

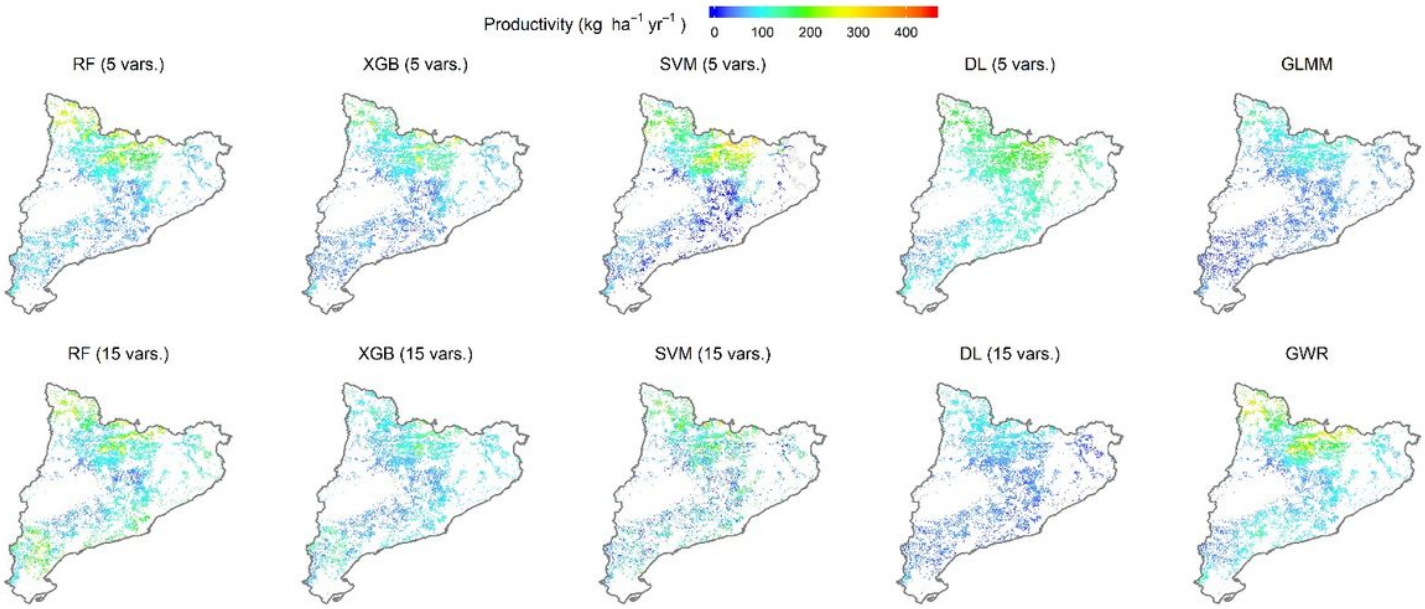


Figure 5

Landscape-level prediction of total annual mushroom productivity, using random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and deep learning (DL) trained with 15 and 5 variables, respectively, and generalized linear mixed models (GLMM) and geographically weighted regression (GWR) fitted using 5 predictors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- floatimage7.png
- floatimage8.png
- floatimage9.png
- Onlinefloatimage10.Png
- Onlinefloatimage11.Png
- Onlinefloatimage12.Png
- Onlinefloatimage13.Png