Anticipating species distributions: handling sampling effort bias under a Bayesian framework

Duccio Rocchini [1,*], Carol X. Garzon-Lopez [2], Matteo Marcantonio [1,3,*], Valerio Amici[4], Giovanni Bacaro[5], Lucy Bastin[6,7], Neil Brummitt[8], Alessandro Chiarucci[9], Giles M. Foody[10], Heidi C. Hauffe[1], Kate S. He[11], Carlo Ricotta[12], Annapaola Rizzoli[1], Roberto Rosà[1]

December 2, 2016

[1] Fondazione Edmund Mach, Research and Innovation Centre, Department of Biodiversity and Molecular Ecology, Via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy, corresponding author: ducciorocchini@gmail.com, duccio.rocchini@fmach.it

[2] UR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, FRE 3498 CNRS), 9 Université de Picardie Jules Verne, 1 rue des Louvels, FR-80037 Amiens Cedex 1, France; 2

[3] Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California, Davis, USA

[4] Department of Life Sciences, Universitfy of Siena, Via P.A. Mattioli 4, 53100 Siena, Italy

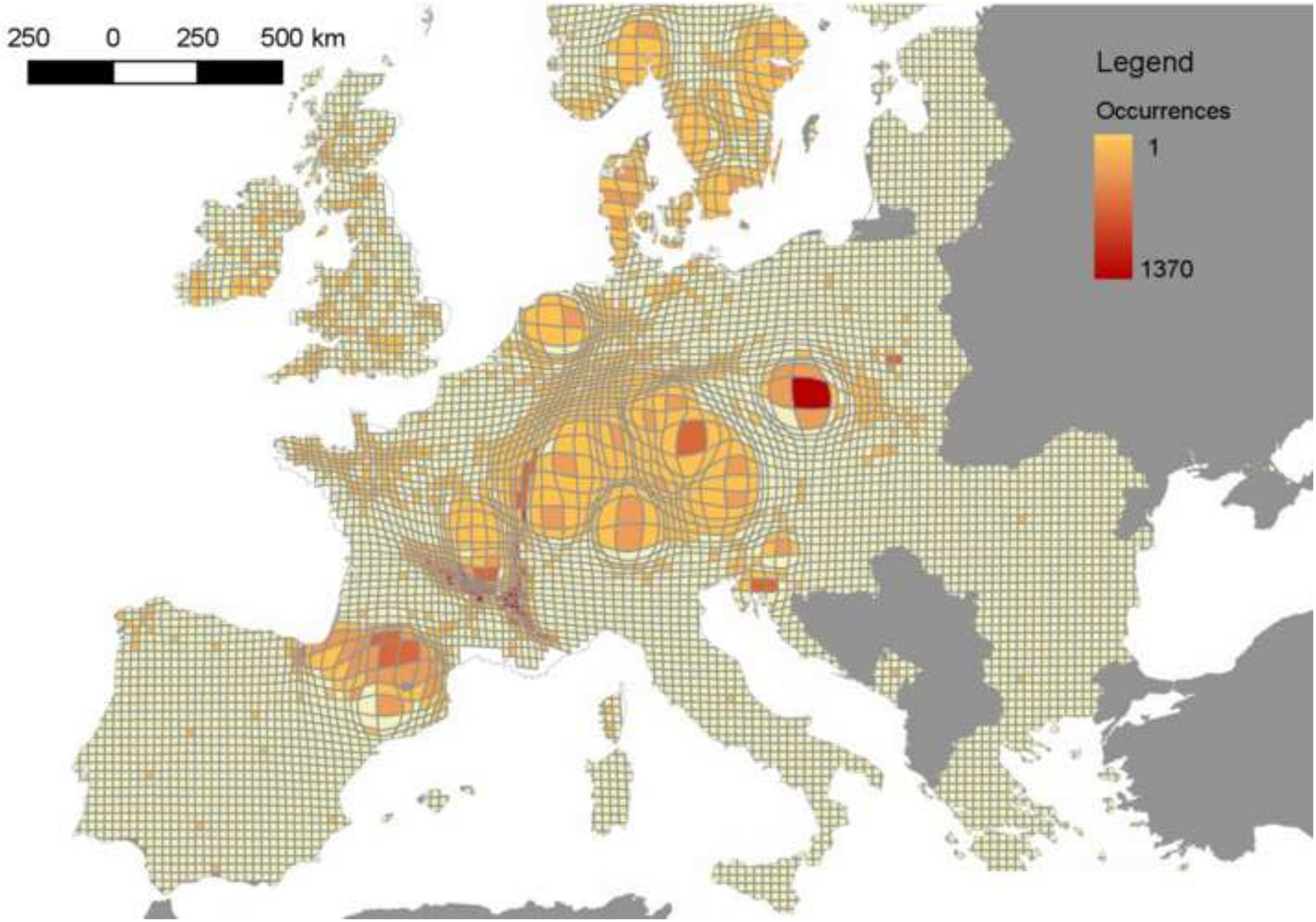
[5] Department of Life Sciences, University of Trieste, Via L. Giorgieri 10, 34127 Trieste, Italy

[6] School of Computer Science, Aston University, UK

[7] Currently on secondment to Joint Research Centre of the European Commission

[8] Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK

***Graphical Abstract**



[9] BIGEA, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum, University of Bologna, Via Irnerio 42, 40126, Bologna, Italy

[10] University of Nottinhgam, University Park, Nottingham, NG7 2RD, UK

[11] Department of Biological Sciences, Murray State University, Murray, Kentucky, 42071, USA

[12] Department of Environmental Biology, University of Rome "La Sapienza", 00185 Rome, Italy

[*] Authors contributed equally to the paper

Abstract

Anticipating species distributions in space and time is necessary for effective biodiversity conservation and for prioritising management interventions. This is especially true when considering invasive species. In such a case, anticipating their spread is important to effectively plan management actions. However, considering uncertainty in the output of species distribution models is critical for correctly interpreting results and avoiding inappropriate decision-making. In particular, when dealing with species inventories, the bias resulting from sampling effort may lead to an over- or under-estimation of the local density of occurrences of a species. In this paper we propose an innovative method to i) map sampling effort bias using cartogram models and ii) explicitly consider such uncertainty in the modeling procedure under a Bayesian framework, which allows the integration of multilevel input data with prior information to improve the anticipation species distributions.

Keywords: anticipation, Bayesian theorem, sampling effort bias, Species Distribution Modeling, uncertainty

List of acronyms: DIC: Deviance Information Criterion, MCMC: Markov Chain Monte Carlo, PPD: posterior probability distribution, SDM: Species Distribution Models

1 Introduction

Anticipation is an important topic in ecological fields such as food science (Lobell et al., 2012), community ecology (Keddy, 1992), species distribution modeling (Willis et al., 2009), landscape ecology (Tattoni et al., in press), and biological invasion science (Rocchini et al., 2015). Anticipatory methods are also crucial for developing effective management practices to deal with invasive species (Rocchini et al., 2015).

Invasive species can modify the structure and functioning of ecosystems, altering biotic interactions and homogenizing previously diverse plant and animal communities over large spatial scales, ultimately resulting in a loss of genetic, species and ecosystem diversity (Winter et al., 2009). The annual economic impact of invasive species has been estimated at over 100 billion dollars just within the USA (NRC, 2002), an order of magnitude higher than those caused by all natural disasters put together (Ricciardi et al., 2011); some authors go as far as to claim that the economic impact of invasive species is incalculable (Mack et al., 2000).

Given the massive negative economic and ecological effects of invasive species, a robust method for predicting species' distributions is crucial for an early assessment of species invasions and effective application of appropriate management actions (Malanson and Walsh, 2013).

Investigating how biodiversity is distributed spatially and temporally across the globe has long been a central theme in ecology (Gaston, 2000) and the methods developed to answer this question have become key tools for biodiversity monitoring (Ferretti and Chiarucci, 2003). For example, species distribution models (SDMs) have been used to map the current distribution of a single species (Rocchini et al., 2011), model the potential distribution of native and invasive species (Rocchini et al., 2015), investigate the statistical performance of different models to infer the distribution of species under various ecological conditions (Guisan and Zimmermann, 2000; Elith and Graham, 2009), test the transferability in space of modeled distribution patterns (Randin et al., 2006; Heikkinen et al., 2012), predict long term changes to species distributions (Pearman et al., 2008) and make inferences on future biodiversity scenarios (Pompe et al., 2008; Engler et al., 2009), evaluate the potential of satellite imagery bands as predictors of biodiversity patterns (Mathys et al., 2009), analyse spatial autocorrelation in species distributions (Carl and Kühn, 2007; Dormann, 2007), and understand biogeographical patterns (Sax, 2001).

In combination with remote sensing products (e.g. Rocchini (2007); Feilhauer et al. (2013)) and current global data sets on in situ species observations, SDMs have become the method of choice for monitoring biodiversity at multiple spatial and temporal scales. However, the strength of this combination depends on the careful selection and application of integrative modeling approaches, in combination with a thorough assessment of uncertainty in both data inputs and modeling methods.

Reliable anticipation of species invasions depends on the quality of input data on one hand and robustness of the predictive SDM on the other. As an example, Rocchini et al. (2011) demonstrated theoretically that input data arising from biased species distribution maps could potentially lead to unsuitable management strategies. In addition, Elith and Leathwick (2009) demonstrated that, given the same input data set, different SDMs might lead to dissimilar results (see also Bierman et al. (2010); Manceur and Kühn (2014)).

The aim of this manuscript is to propose coherent and straightforward methods to explicitly account for uncertainty when mapping species distributions in the light of anticipating the spread of invasive species. In particular we will cover: i) explicitly mapping uncertainty in sampling bias, ii) mitigating uncertainty in data through prior beliefs and Bayesian inference and iii) reporting uncertainty in species distribution maps through Markov Chain Monte Carlo methods. The findings of this manuscript should be of particular interest to landscape managers and planners attempting to predict the spread of species and deal with errors in species distribution maps in a straightforward manner.

2 Mapping input uncertainty related to sampling effort bias

In anticipating species distributions a first step is to ensure that the information indicating where species are present is bias-free or, at least, that the uncertainty of input data is explicitly taken into account in further modeling steps.

One of the main problems with field data on species distributions is related to "sampling effort bias" (Rocchini et al., 2011), namely the bias inherent in some areas being under-sampled with respect to others. Quantifying and mapping the uncertainty derived from variation in the number of observations due to sampling effort can be achieved using cartograms (Gastner and Newman, 2004), in which the shape of spatial objects (e.g. polygons, cells, etc.) is directly related to a determined property, in our case to uncertainty.

Cartograms build on the standard treatment of diffusion theory by Gastner and Newman (2004), in which the current spatial density of a population is given by:

$$J = v(r,t)p(r,t) \tag{1}$$

where v(r, t) and p(r, t) are the velocity and density of the spread of the population under study, respectively, at position r and time t.

Cartograms facilitate the visualization of spatial uncertainty in the data by varying the size of each polygon according to the density of information contained (e.g. number of observations, variation, etc.). As an example, we show a cartogram of the distribution of *Abies alba* Miller overlapping a grid to the set of records obtained from the Global Biodiversity Information Facility (GBIF, http://www.gbif.org, Figure 1). GBIF offers free and open access to hundreds of millions of records from over 30,000 species datasets which are collated from around the world and stored with a common Darwin Core data standard. The cartogram was developed using the free and open source software ScapeToad (http://scapetoad.choros.ch/). Since cells with a higher number species occurrences might be biased by the effort spent visiting them, in Figure 1, the shape of each cell is determined by the the number of times it was visited (i.e. number of different dates recorded in GBIF for the species in that cell). From now on, we will refer to this as sampling effort. The colour represents the spatial distribution (density of occurrences, sensu) of the species in each cell.

Therefore, cartograms allow uncertainty to be shown explicitly in a straightforward manner. Furthermore, sampling effort might be considered as a variable in the SDM procedure, as described in the next section.

3 Accounting for input uncertainty in the modeling procedure: multi-level models, prior beliefs and probability distribution surfaces

Species observation records are often heterogeneous and incomplete because, for example, they are unevenly distributed by year or area, or were collected by different field operators. In addition, there is wide variation in recording behaviours.

GBIF is a classic example of such heterogeneity: GBIF data is opportunistically gathered from a mixture of systematic surveys and volunteer projects, and the intensity of publishing effort is strongly influenced by the membership of the organisation. In terms of geographic coverage, GBIF contains plentiful data from Northern Europe and America, parts of Latin and Central America, South Africa, Australia and Oceania – but by contrast, there are significant gaps in other regions, and there is a large variation in sampling effort even between neighbouring European countries (see Appendix 1, Figure S1). This heterogeneity makes it difficult to estimate the underlying variable (actual species presence and density of occurrences) and potentially has an enormous impact on the information content of any one species observation or set of observations (Isaac and Pocock, 2015). This paper proposes methods by which ancillary knowledge about a species and its environment might be exploited in a Bayesian framework to increase that information content.

Multi-level models can be essential for detecting (spatially) clustered data by considering the variation between groups (clusters). This approach is more efficient and powerful than standard linear modeling techniques as it provides a coherent and flexible method for modeling the effects of sampling variation and allows uncertainty to be elegantly accounted for at all levels of data structure (Gelman and Hill, 2006).

Furthermore, environmental variables with different spatial or temporal resolution (i.e., country, regional or pixel level) are often used as predictors in SDMs. Multi-levels models can simultaneously and coherently incorporate multi-level predictors allowing effects to be modelled at the appropriate scale (Gelman and Hill, 2006). Hierarchical models are naturally handled using Bayesian methods, which provide intuitive and direct estimates of uncertainty around parameter estimates (Link et Sauer, 2002).

Despite tremendous effort by ecologists, collecting unbiased and reliable data on the presence of species in a determined area/time to assess their potential distribution through SDMs is sometimes not feasible since systematic field work is inherently expensive, time-consuming, and often involves logistical hurdles, if the species under study is, for example, rare, elusive, inhabits remote areas, or is in transitional equilibrium with its ecological niche (as is the case with invasive species). Even for less problematic species, presence/absence data may also be distorted by several potential flaws, such as sampling errors and subjectivity. As a result, SDM outputs may show high uncertainty and be difficult to interpret, jeopardizing their utility in conservation applications. However, besides the availability of observation data directly exploitable for modeling purposes, there is a wider set of ecological data that can be used in SDMs, the so called "prior knowledge". This data is very often neglected and comprises information represented in different formats; for example, previously conducted experiments, scientific literature on the studied species or similar species, or even as "prior beliefs" (basic ecological principles). Bayesian inference allows basic ecological principles and prior data to be incorporated in a straightforward manner with potential cost-effective consequences in increasing confidence of SDMs (McCarthy and Masters, 2005; Bierman et al., 2010; Manceur and Kühn, 2014). The prior information needs to be translated into a probability distribution, which is then combined under Bayes' rule with the likelihood information contained in the original data to estimate a "posterior belief" or posterior probability distribution (PPD). The contribution of the prior and the data to the posterior distribution depends on their relative precision, with the more precise of the two having the greatest effect. A prior distribution can be non-informative (flat prior), mildly informative (vague prior) or informative (strong prior). In any case, the prior must be clearly described and justified according to the context under investigation (Kruschke, 2015).

The result of the interaction between the likelihood of the data and the prior distribution is itself a probability distribution (posterior probability distribution or PPD). In an SDM, the advantage of having model parameter estimations expressed as probability distributions, and not as point estimation of the mean, is that the predicted suitability of the species in each prediction unit (pixel) is itself a probability distribution. The suitability of the PPD in each spatial unit represents the uncertainty of the prediction in that unit. This uncertainty is stored in the Markov Chain Monte Carlo (MCMC) model and can be re-used in future modeling exercises that, for example, use a different set of data.

As an example, we applied a multi-level logistic regression with Bayesian inference to model the distribution of *Abies alba* in Europe. We chose this species due to its well known autoecology and actual distribution in Europe (Farjon, 1998; Tinner et al., 2013; Gazol et al., 2015). We derived 44375 *Abies alba* presence records from the GBIF database, as points in vector format (see Appendix 1, Figures S3 and S4). We generated an equal number of pseudoabsences using the following strategy: we selected random points a) within areas where conifers have been sampled (conifer occurrences in the GBIF dataset) to pick the same areas that have been surveyed using the sampling protocol used to record *Abies alba* presences, b) outside dry climatic zones (e.g. Mediterranean climate) derived from the Köppen-Geiger climatic zones map (Köppen and Geiger, 1930) where this species is not found and c) outside a radius of 100 metres around the presence points to avoid overlap with presence points.

We generated an equal number of absence locations at areas within which conifers have been sampled (conifer occurrences in the GBIF dataset) and outside a 100 meters radius from the presence points and the temperate and dry climatic zones (e.g. mediterranean climate) derived from the Köppen-Geiger climatic zones map .

To select the predictor variables, we performed a literature review on the ecology of the species (Aussenac, 2002; Wolf, 2003; Rolland et al., 2009; Tinner et al., 2013; Gazol et al., 2015). Hence, we relied on three different datasets by selecting: i) the annual mean temperature (Bio1), and mean diurnal temperature range (Bio2) obtained from the WorldClim dataset (Hijmans et al., 2005), ii) radiation seasonality (Bio23) and the annual mean moisture index (Bio28), obtained from the CliMond dataset (Kriticos et al., 2012), and iii) the number of wet days during summer and frost days during winter (and early spring) derived from the wet-days and ground-frost data in the climate research unit dataset (Mitchell et al., 2004) (see Figure 2).

Considering sampling effort as a predictor, the sampling of the GBIF dataset is clearly opportunistic. As a result, the unevenness of sampling effort is particularly evident, with the Northern European region being more sampled than other European regions (see Appendix 1, Figure S1). This bias in GBIF data could generate unreliable predictions.

The clustering of GBIF data mainly derives from differences in surveys at national and subnational level (Appendix 1, Figure S1). Thus, the sampling effort was derived as the number (richness) of dates of survey recorded in the GBIF dataset per polygon of the official administrative division of European countries using the Nomenclature of Territorial Units for Statistics level 3 (NUTS 3).

We built a multi-level model to take into account the different resolution of the predictor variables (Figure 2) and the differential sampling effort of *Abies alba* occurrences in each NUTS3 polygon. The sampling effort was used to re-scale the precision of the likelihood at pixel level, multiplying the scaled sampling effort by the standard deviation of the Gaussian likelihood. As a result, the likelihood estimate of pixels in regions with a higher number of samples was expected to be more precise. The theoretical model (Figure 2) was coded in JAGS language and run in JAGS 4.2.0 through R (R Core Team, 2016) using the R2jags (Su and Yajima, 2016) and CODA (Plummer et al., 2002) packages. In order to allow reproducibility (Rocchini and Neteler (2012)) of our approach we have included the complete R code in Appendix 2.

As previously stated, in heterogeneous datasets like the GBIF set, the sampling effort in a certain region may be correlated with the presence of the species under study. Therefore, a more highly sampled region should have also a higher probability of hosting the species. However, our data showed a weak sampling effort signal, with a high number of very low-sampled regions showing presence of *Abies alba*. This may result from errors, or low numbers of records not being representative of the distribution of the species under study. Therefore, we applied uninformative priors ($\mu = 0$, SD = $^1/_{10^{-2}}$) for all the predictors but not for sampling effort, whose prior distribution $p(\theta)$ was given three different sets of parameters:

$$p(\theta) = \begin{cases} dnorm(0, {}^1/_{10^{-2}}), & \text{uninformative prior.} \\ dnorm(1, 10), & \text{mild positive prior.} \\ dnorm(5, 5), & \text{strong positive prior.} \end{cases} \tag{2}$$

Such distributions were chosen as examples under the hypothesis that *i)* data alone were enough to account for heterogeneity in sampling effort; *ii)* a mildly informative (vague) prior knowledge about the positive correlation of sampling effort was useful for improving the model; *iii)* imposing strong prior knowledge on the positive influence of the prior would improve the model output. These three hyphoteses were translated in three models that shared the same structure (Figure 2) exept for the prior distribution imposed on sampling effort. All the predictors were scaled and centered in order to improve the efficiency of the MCMC process. PPDs for all parameters were sampled from each of two chains with 10000 MCMC iterations using 1000 burn-in and 1000 adaptation iterations, with a thinning set of 20. Convergence was assesed by the Gelman-Rubin statistic (Gelman and Rubin, 1992). Each model was then used to estimate the suitability PPDs in each pixel of the study area. The parameter estimates for the three models will show if different prior belief on the role of sampling effort changed the model parameter estimates. Furthermore, the Deviance Information Criterion (DIC, see Spiegelhalter et al. (2014)) was used to assess the model with the best predictive power.

The Posterior Probability Distributions (PPDs) of model parameters for the three models (with different priors on sampling effort, see Equation 2) are reported in Figure 3. All the models agreed on the direction and effect size of the predictors (Figure 3). Credible effects (no intersection with 0 in Figure 3) were attained for those variables directly related to temperature. In particular, annual mean temperature (Bio1 and Bio1^2) and radiation seasonality (Bio23) showed negative effects while mean diurnal temperature range (Bio2) showed positive effects (Figures 3 and 4). The negative credible effect of Bio1^2 implies that the relationship between the probability of presence (suitability) of *Abies alba* and annual mean temperature has a “bell shape”, by rising slowly to the left of the annual mean temperature average (7.8 °C) and decreasing rapidly when on its right (Figure 4). On the contrary, the distribution of wet days, annual mean moisture index (Bio28) and frost days included 0, showing a non-credible effect on the presence of *Abies alba*.

The sampling effort coefficient changed heavily between models. In the first model with an uninformative prior, the coefficient average was slightly negative but with its high density interval comprising 0 (Figure 3). Therefore we concluded that according to the data the sampling effort had a non-credible effect. In the second model (Figure 3) a mildly informative positive prior affected the estimate of the parameters, but yet was not enough to derive a credible effect of the prior estimate. In the last model, the strong informative prior pulled the estimation of sampling effort coefficient towards positive values. This showed that, according to the data and to the “prior knowledge", the sampling effort was positively affecting the probability of presence of *Abies alba.*

In summary, the model with the strong prior showed an improved precision of sampling effort, basically maintaining that of the others (Figure 3). Based on this and since the DIC did not show differences for the strong prior-model with respect to the uninformative prior-model (Table 1, $\delta DIC \leq 4$, see Burnham and Anderson (2002)), we further focused on the model with a strong prior to build the output distribution map. The resulting potential niche distribution of *Abies alba* is thus shown in Figure 5.

4 Discussion

In this paper, we have demonstrated the importance of i) mapping uncertainty derived from varying sampling effort and ii) considering it in an explicit manner in order to anticipate species' potential distributions. We have provided a case study with a plant species widespread throughout Europe (*Abies alba*) where the observed data (Figure 1) and the modelled potential niche (Figure 5) differed mainly because of tree plantations recorded in the GBIF dataset. For example, Northern Europe was shown to be unsuitable for the natural spread of the species in our Bayesian model (Figure 5), as well as in previous studies on the distribution of the species (e.g. the European Forest genetic Resources programme, http://www.euforgen.org/, see Appendix 1, Figure S2), corroborating our results. However, it appeared to be present in the GBIF field-based dataset (Figure 1, see also Appendix 1, Figure S3), mainly because of human-related conifer plantations.

Notably, when we associated a stronger prior to sampling effort, model coefficient estimates had lower uncertainty, and in addition, the model DIC did not differ from the model with the uninformative prior. Therefore, a strong prior allowed us to decrease uncertainty and maintain high model quality ($\delta DIC \leq 4$, see Burnham and Anderson (2002)).

We have shown that multilevel models coupled with Bayesian inference can be used to account for variability in sampling effort, integrating external data on prior knowledge with species observations, to model species distribution more accurately and with higher certainty than previous methods. The priors considered in the reported case study were only examples generated here to illustrate how the precision of parameter estimates can potentially be increased using prior knowledge about the system under study. However, in order to have scientifically sound results, the priors considered should obviously be fully justified and rooted in ecological theory.

Anticipating species potential distributions based on prior information

(Bayesian modeling) can help to predict the potential future spread of a species in space (and time) in a robust manner (Bierman et al., 2010; Manceur and Kühn, 2014). Using sampling effort bias among priors was important in our case since it allowed such uncertainty to be considered explicitly in the model. This can help to accommodate the error rate directly into the modeling procedure.

Hence, calibrating models conditioned on previous knowledge and/or observations might be feasible when relying on a Bayesian framework in which:

$$P(Y|H) \tag{3}$$

where P = the probability of occurrence of patterns Y given a hypothesis H is substituted by:

$$P(H|Y) \tag{4}$$

i.e. the probability P that a hypothesis H is true in light of the available data.

Bayesian statistics have long been used in independent scientific disciplines and topics such as trait loci mapping (Ball, 2001), environmental science (Clark, 2005), machine learning approaches in computer science (Dietterich, 2000), classification of remotely-sensed images (Goncalves et al., 2009), conservation genetics (Bertorelle et al., 2004), statistical algorithm development (Hoeting et al., 2009) and sampling strategies (Mara et al., 2016).

In the framework of ecological patterns and processes, Ellison (2004) makes an explicit quest for using known information to build a model, relying on prior rather than posterior probabilities. This reinforces the view of Ginzburg et al. (2007) that biology should constrain mathematical constructions. Quoting the authors, "While mathematics provides an incredibly vast set of possible equations, logic dictates that only a small subset of these equations can represent a given ecological phenomenon. A large number of constructions, while mathematically sound, should be excluded based on their inconsistency with biology."

This is especially true when the results of model construction impact decision-making, which could be more focused and effective if uncertainty was explicitly taken into account based on previous literature regarding the main drivers that shape the distribution of species (Ellison, 1996). Our approach reduces the danger of relying on misleading predictions of alien species invasions with high model errors, which are hidden or unrecognizable using previous approaches (Rocchini et al., 2015).

In the framework of Species Distribution Modeling it has been demonstrated that prior probabilities in the observation of a certain species might improve model performance. This is true at various hierarchical levels, from species to entire communities. Thus, applying Bayes' theorem to predict values at a certain site might thus allow known environmental properties to be accounted for. If Bayesian models do not outperform other modeling techniques, they at least better reflect the theory under the realized niche of a certain species. A number of examples are provided in Guisan and Zimmermann (2000), modeling different plant species in different habitat types.

5 Conclusion

In the light of the importance of anticipating species future distributions, especially for economically important invasive species, it is crucial to detect those areas into which such a species might be expected to disperse. Anticipating their spread based on the suitability of environmental conditions can lead to more effective management strategies, allowing timely actions to be initiated and preventing further spread (Rocchini et al., 2015).

This can be summarized by the following equation:

$$Decision = \begin{pmatrix} < E_m| > I & < E_m| < I \\ > E_m| > I & > E_m| < I \end{pmatrix} \tag{5}$$

In this case, a high (or low) invasion rate I might be related to high or low error E_m in the output model being observed by decision makers. The most dangerous situation is when a low predicted invasion rate is related to a high error in the modeling procedure. In this case decision makers might underestimate the effort against the likelihood of invasion, that, from the species distribution map, is suspected to be low.

In this paper we have demonstrated the power of incorporating sampling bias into the model being used by relying on prior probabilities of distribution of a plant species widely spread in Europe. We believe this is a good example to further encourage species distribution modellers and environmental planners and conservationists to account for uncertainty and bias in the sampling effort in anticipating the spatial spread of species, instead of relying on distribution maps with potentially hidden uncertainty.

6 Acknowledgments

We are particularly grateful to the handling Editor Rocco Scolozzi and to two anonymous Reviewers who provided useful insights which improved a first draft of this manuscript. We thank Ingolf Kühn for precious suggestions.

DR was partially supported by the EU BON (Building the European Biodiversity Observation Network) project, funded by the European Union under the 7th Framework programme (Contract No. 308454), by the ERANET BioDiversa FP7 project DIARS, funded by the European Union and by the Life project Future For CoppiceS.

References

Alba-Sánchez, F., López-Sáez, J.A., Pando, B.B., Linares, J.C., Nieto-Lugilde, D., López-Merino, L., 2010. Past and present potential distribution of the Iberian Abies species: a phytogeographic approach using fossil pollen data and species distribution models. Diversity and Distributions 16, 214–228. doi:10.1111/j.1472-4642.2010.00636.x

Aussenac, G. (2002). Ecology and ecophysiology of circum-Mediterranean firs in the context of climate change. Annals of Forest Science 59: 823-832.

Ball, RD. (2001). Bayesian Methods for Quantitative Trait Loci Mapping Based on Model Selection: Approximate Analysis Using the Bayesian Information Criterion." Genetics, 159: 1351-1364.

Barbosa, A., 2015. fuzzySim: applying fuzzy logic to binary similarity indices in ecology. Methods in Ecology and Evolution p. in press.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M., Dormann, C.F. (2012). What's on the Horizon for Macroecology? Ecography, 35: 673-683.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics, 19: 10-15.

Bertorelle, G., Bruford, M., Chemini, C., Vernesi, C., Hauffe, H.C. (2004). New, flexible Bayesian approaches to revolutionize conservation genetics. Conservation Biology, 18: 1-2.

Bierman, S.M., Butler, A., Marion, G., Kühn, I. (2010). Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. Ecography, 33: 451-460.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A practical Information-Theoretic Approach, Second. ed. Springer.

Carl, G., Kühn, I. (2007). Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. Ecological Modelling, 207: 159-170.

Chen, G., Kéry, M., Plattner, M., Ma, K., Gardner, B. (2012). Imperfect detection is the rule rather than the exception in plant distribution studies. Journal of Ecology, 101: 183-191.

Clark, J. (2005) Why environmental scientists are becoming Bayesians. Ecology Letters, 8: 2-14.

Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. Lecture Notes in Computer Science, 1857: 1–15.

Dormann, C.F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology and Biogeography, 16: 129-138.

Elith, J., Graham, C.H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography, 32: 66-77.

Elith, J., Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40: 677-697.

Ellison, A. (1996). An Introduction to Bayesian Inference for Ecological Research and Environmental Decision-Making. Ecological Applications 6: 1036-1046.

Ellison, A. (2004). Bayesian Inference in Ecology. Ecology Letters, 7: 509-520.

Engler, R., Randin, C.F., Vittoz, P., Czaka, T., Beniston, M., Zimmermann, N.E., Guisan, A. (2009). Predicting future distributions of mountain plants under climate change: does dispersal capacity matter? Ecography, 32: 34-45.

Farjon (1998), World Bibliography and Checklist of Conifers, RBG Kew.

Feilhauer, H., Thonfeld, F., Faude, U., He, K.S., Rocchini, D., Schmidtlein, S. (2013). Assessing floristic composition with multispectral sensors - A comparison based on monotemporal and multiseasonal field spectra. International Journal of Applied Earth Observation and Geoinformation, 21: 218-229.

Ferretti, M., Chiarucci, A. (2003). Design concepts adopted in long-term forest monitoring programs in Europe - problems for the future? Science of The Total Environment, 310, 171-178.

Gaston, K.J. (2000). Global patterns in biodiversity. Nature 405, 220-227.

Gastner, M.T., Newman, M.E.J. (2004). Diffusion-based method for producing density-equalizing maps. Proceedings of the national Academy of Sciences USA, 101: 7499-7504.

Gazol, A., Camarero, J.J., Gutierrez, E., Popa, I., Andreu-Hayles, L., Motta, R., Nola, P., Ribas, M., Sangüesa-Barreda, G., Urbinati, C., Carrer, M. (2015). Distinct effects of climate warming on populations of silver fir (*Abies alba*) across Europe. Journal of Biogeography, 42: 1150-1162.

Gelman, A., Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7: 457-511.

Gelman, A., Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Ginzburg, L.R., Jensen, C.X.J., Yule, J.V. (2007). Aiming the "unreasonable effectiveness of mathematics" at ecological theory, Ecological Modelling, 207, 356-362.

Goncalves, Luisa, Cidalia Fonte, Eduardo Julio, and Mario Caetano. "A Method to Incorporate Uncertainty in the Classification of Remote Sensing Images." International Journal of Remote Sensing 30, no. 20 (2009): 5503, 5489.

Guisan, A., Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. Ecological Modelling, 135: 147-186.

Heikkinen, R.K., Marmion, M., Luoto, M. (2012). Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35: 276-288.

Hijmans, R..J, Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25: 1965-1978.

Hoeting, JA, CT Volinsky, and D Madigan. “Bayesian Model Averaging: A Tutorial.” Statistical Science 14, no. 4 (1999): 417, 382.

Isaac, N.J.B., Pocock, M.J.O. (2015). Bias and information in biological records. Biological Journal of the Linnean Society, 115: 522-531.

Keddy, P.A. (1992). Assembly and response rules: two goals for predictive community ecology. Journal of Vegetation Science 3: 157-164.

Köppen, W., Geiger, R. (1930). Handbuch der klimatologie. Gebrüder Borntraeger Berlin, Germany.

Kriticos, D.J., Webber, B.L., Leriche, A., Ota, N., Bathols, J., Macadam, I., Scott, J.K. (2012). CliMond: global high resolution historical and future scenario climate surfaces for bioclimatic modelling. Methods in Ecology and Evolution, 3: 53-64.

Kruschke, J.K., 2015. Doing Bayesian Data Analysis, 2nd Edition, Elsevier, Amsterdam.

Link, W.A., Sauer, J.R., 2002. A hierarchical analysis of population change with application to Cerulean Warblers. Ecology 83, 9.

Lobell, D. B., Sibley, A., Ortiz-Monasterio, J.I. (2012). Extreme heat effects on wheat senescence in India. Nature Climate Change, 2: 186-189.

Mack, R.N., Simberloff, D., Lonsdale, W.M., et al. (2000). Biotic invasions: causes, epidemiology, global consequences, and control. Ecological Applications, 10: 689-710.

Manceur, A.M., Kühn, I. (2014). Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. Methods in Ecology and Evolution, 5: 739-750.

Mathys, L., Guisan, A., Kellenberger, T.W., Zimmermann, N.E. (2009). Evaluating effects of spectral training data distribution on continuous field mapping performance. ISPRS Journal of Photogrammetry and Remote Sensing, 64: 665-673.

Malanson, G.P., Walsh, S.J. (2013). A geographical approach to optimization of response to invasive species. In: Walsh SJ and Mena C (eds) Science and Conservation in the Galapagos Islands: Frameworks and Perspectives. New York, USA: Springer, pp. 199-215.

Mara, T.A., Delay, F., Lehmann, F., Younes, A. (2016). A Comparison of Two Bayesian Approaches for Uncertainty Quantification. Environmental Modelling & Software, 82: 21-30.

McCarthy, M.A., Masters, P., 2005. Profiting from prior information in Bayesian analyses of ecological data. Journal of Applied Ecology 42, 1012–1019. doi:10.1111/j.1365-2664.2005.01101.x

Mitchell, T.D., Carter, T.R., Jones, P.D., Hulme, M., New, M. (2004). A comprehensive set of climate scenarios for Europe and the globe: the observed record (1900-2000) and 16 scenarios (2000-2100). University of East Anglia, Norwich,UK, pp. 30.

NRC (Committee on the Scientific Basis for Predicting the Invasive Potential of Non-indigenous Plants and Plant Pests in the United States) (2002) Predicting Invasions of Non-indigenous Plants and Plant Pests. Washington, DC: National Academy Press.

Pearman, P.B., Randin, C.F., Broennimann, O., Vittoz, P:, van der Knaap, W.O., Engler, R., Le Lay, G., Zimmermann, N.E., Guisan; A. (2008). Prediction of plant species distributions across six millennia. Ecology Letters, 11: 357-369.

Pompe, S., Hanspach, J., Badeck, F., Klotz, S., Thuiller, W., Kühn, I. (2008). Climate and land use change impacts on plant distributions in Germany. Biology Letters, 4: 564-567.

Plummer, M., Best, N., Cowles, K., Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 6: 7-11.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A. (2006). Are niche-based species distribution models transferable in space? Journal of Biogeography, 33: 1689-1703.

Ricciardi, A., Palmer M.E., Yan, N.D. (2011) Should biological invasions be managed as natural disasters? Bioscience, 61: 312-317.

Rocchini, D. (2007). Distance decay in spectral space in analysing ecosystem b-diversity. International Journal of Remote Sensing, 28: 2635-2644.

Rocchini, D., Andreo, V., Förster, M., Garzon-Lopez, C.X., Gutierrez, A.P., Gillespie ,T.W., Hauffe, H.C., He, K.S., Kleinschmit, B., Mairota, P., Marcantonio, M., Metz, M., Nagendra, H., Pareeth, S., Ponti, L., Ricotta, C., Rizzoli, A., Schaab, G., Zebisch, M., Zorer, R., Neteler, M. (2015). Potential of remote sensing to predict species invasions - a modeling perspective. Progress in Physical Geography, 39: 283-309.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography, 35: 211-226.

Rocchini, D., Neteler, M. (2012). Let the four freedoms paradigm apply to ecology. Trends in Ecology & Evolution, 27: 310–311.

Rolland, C., Michalet, R., Desplanque, C., Petetin, A., Aimé, S. 2009. Ecological requirements of Abies alba in the French Alps derived from dendro-ecological analysis. Journal of Vegetation Science, 10: 297-306.

Sax, D.F. (2001). Latitudinal gradients and geographic ranges of exotic species: implications for biogeography. Journal of Biogeography, 28: 139-150.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2014). The deviance information criterion: 12 years on (with discussion). Journal of the Royal Statistical Society, Series B, 76: 485-493.

Su, Y.-S., Yajima, M. (2016). R2jags: A Package for Running jags from R. http://CRAN. R-project. org/package= R2jags

Tattoni, C., Ianni, E., Geneletti, D., Zatelli, P., Ciolli, M. (in press). Landscape changes, traditional ecological knowledge and future scenarios in the Alps: A holistic ecological approach. Science of the Total Environment.

Tinner, W., Colombaroli, D., Heiri, O., Henne, P.D., Steinacher, M., Untenecker, J., Vescovi, E., Allen, J.R.M., Carraro, G., Conedera, M., Joos, F., Lotter, A.F., Luterbacher, J., Samartin, S., Valsecchi, V. (2013). The past ecology of *Abies alba* provides new perspectives on future responses of silver fir forests to global warming. Ecological Monographs, 83: 419-439.

Willis, S.G., Thomas, C., Hill, J.K., Collingham, Y., Telfer, M.G., Fox, R., Huntley, B. (2009). Dynamic distribution modelling: predicting the present from the past. Ecography 32: 5-12.

Winter, M., Schweiger, O., Klotz, S., Nentwig, W., Andriopoulos, P., Arianoutsou, M., Basnou, C., Delipetrou, P., Didziulis, V., Hejda, M., Hulme, P.E., Lambdon, J.P., Pysek, P., Royl, D.B., Ingolf Kühn (2009). Plant extinctions and introductions lead to phylogenetic and taxonomic homogenization of the European flora. Proceedings of the National Academy of Sciences of the United States of America, 106: 21721-21725.

Wolf, H. (2003). EUFORGEN Technical Guidelines for genetic conservation and use for silver fir (*Abies alba*). International Plant Genetic Resources Institute, Rome, Italy. 6 pp.

638 # Tables

Model	DIC	Gelman diagnostic	Burn In	Iterations	Chains
Uninformative prior	1938	1.13	2000	10000	2
Mild prior	2133	1.15	2000	10000	2
Strong prior	1940	1.22	2000	10000	2

Table 1: Deviance Information Criterion (DIC) used to assess the prior with the best predictive power. Notice that $\delta DIC \leq 4$ using an uninformative prior and a strong prior on sampling effort. Therefore, a strong prior allowed us to decrease uncertainty and maintain high model quality. Refer to the main text for additional information.

639 # Figures

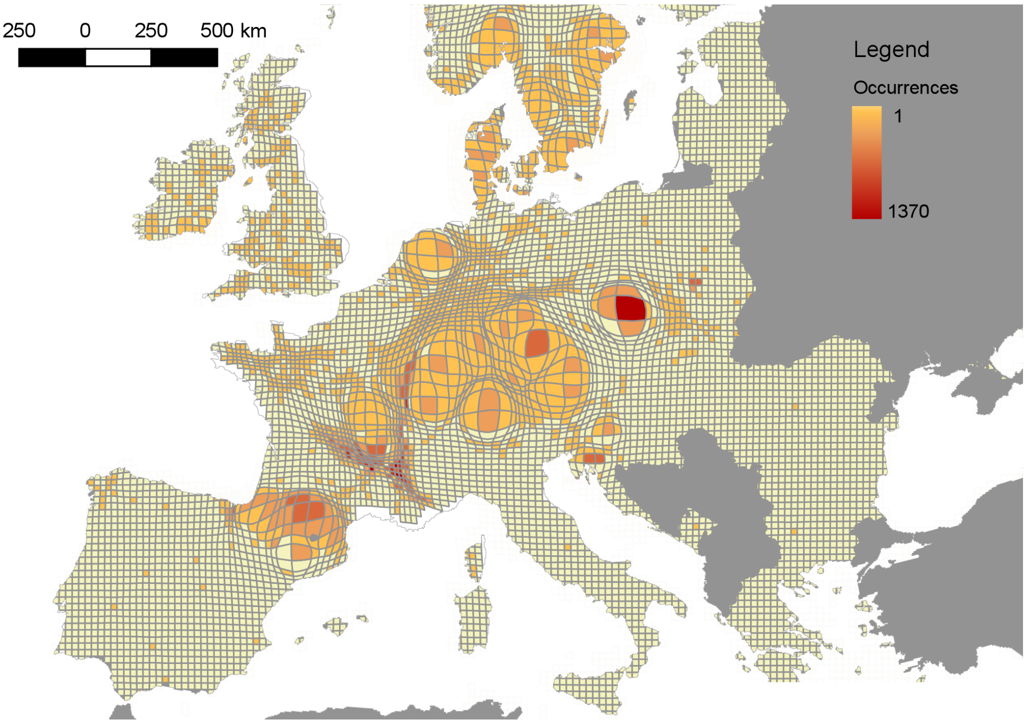
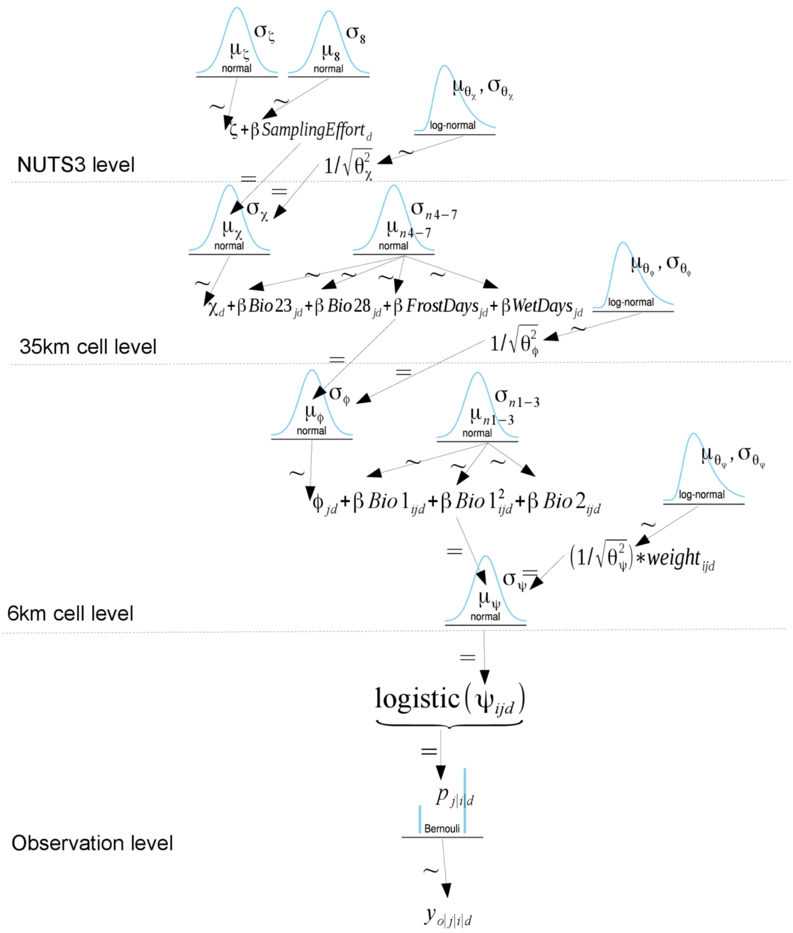


Figure 1: Cartogram representing the sampling effort bias (cell distortion) of the GBIF dataset related to *Abies alba*. This species is not native in Northern Europe, although it is widely cultivated as a timber tree, as thus present in the GBIF dataset.



σ_ζ
μ_ζ
normal
σ_8
μ_8
normal
$\mu_{\theta_\chi}, \sigma_{\theta_\chi}$
log-normal
~
~
$\zeta + \beta\, SamplingEffort_d$
~
NUTS3 level
$1/\sqrt{\theta_\chi^2}$
=
=
σ_χ
μ_χ
normal
σ_{n4-7}
μ_{n4-7}
normal
$\mu_{\theta_\varphi}, \sigma_{\theta_\varphi}$
log-normal
~
~
~
~
~
$\chi_d + \beta\, Bio23_{jd} + \beta\, Bio28_{jd} + \beta\, FrostDays_{jd} + \beta\, WetDays_{jd}$
~
35km cell level
$1/\sqrt{\theta_\varphi^2}$
=
=
σ_φ
μ_φ
normal
σ_{n1-3}
μ_{n1-3}
normal
$\mu_{\theta_\psi}, \sigma_{\theta_\psi}$
log-normal
~
~
~
~
$\phi_{jd} + \beta\, Bio1_{ijd} + \beta\, Bio1^2_{ijd} + \beta\, Bio2_{ijd}$
~
$(1/\sqrt{\theta_\psi^2}) * weight_{ijd}$
=
=
σ_ψ
μ_ψ
normal
6km cell level
=
$\underbrace{\text{logistic}(\psi_{ijd})}$
=
$p_{j|i|d}$
Bernoulli
Observation level
~
$y_{o|j|i|d}$

Figure 2: The multi-level model represented through a pictogram. To select the predictor variables, we performed a literature review on the ecology of the species, finally selecting: radiation seasonality (Bio23), the annual mean moisture index (Bio28), the number of wet days during summer and the frost days during winter and early spring, the annual mean temperature (Bio1), the mean diurnal temperature range (Bio2). Sampling effort was calculated as the richness of dates of survey recorded in the GBIF dataset for each NUTS3 country. Refer to the main text for additional information on the source of each dataset. Symbols used in this figure: μ, σ = mean and standard deviation of prior and hyperprior distributions; ζ, χ, ϕ = intercepts for NUTS3, 35km, 6km level of the model; subscript *d,j,i,o* = index for NUTS3, 35km, 6km and observation level; $weight_{ijd}$ = scaled weights for sampling effort; logistic(ψ) = logistic transformation of the model output (link function); $p_{i|j|d}$ = probability of occurrence; $y_{o|i|j|d}$ = presence or absence. Refer to Kruschke (2015) for a complete dissertation about the terms and the graphical representation of the proposed model. Notice that variables at 6km resolution were resampled from an original resolution of 1km to allow the Bayesian model to be run in R. The R code of the model is available in Appendix 2.

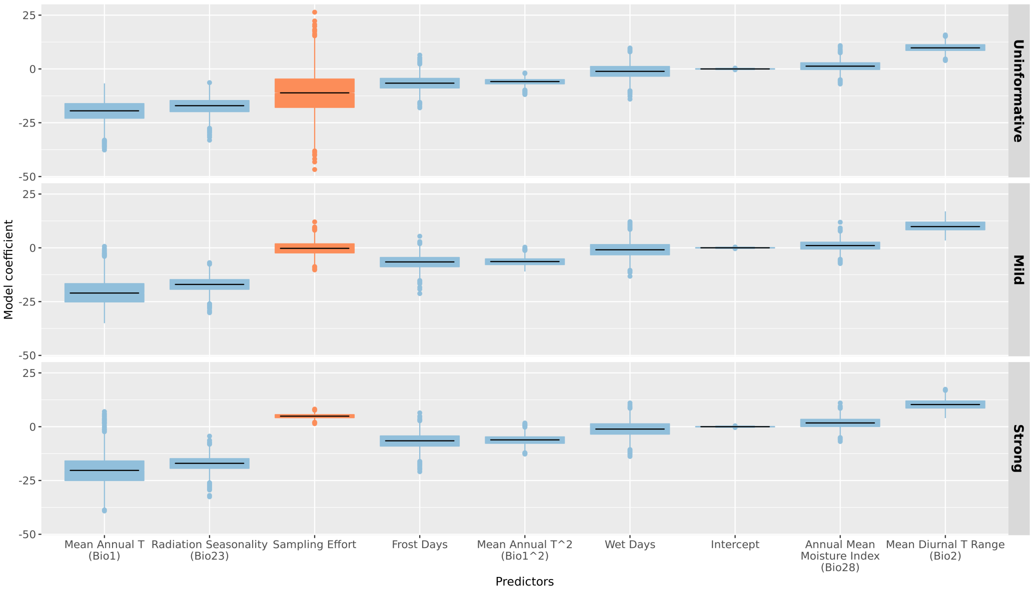


Figure 3: Boxplots of the β coefficient PPDs for the three models (in the three figure facets). Each box represents the 1st and 3rd quartiles of a coefficient distribution, the black horizontal line the distribution median, the whiskers the limits of the 1.5*interquartile range, while the filled circles represent the outlying points. If whiskers did not overlap 0 we inferred as “credible effect”. We showed in red the boxplots reporting the distribution of the β coefficient of the sampling effort. It is clear that the major difference among models was related to the precision of sampling effort, which increased passing from the model with an uninformative prior on sampling effort, through that with a mild prior, reaching its highest value in the model with a strong prior.

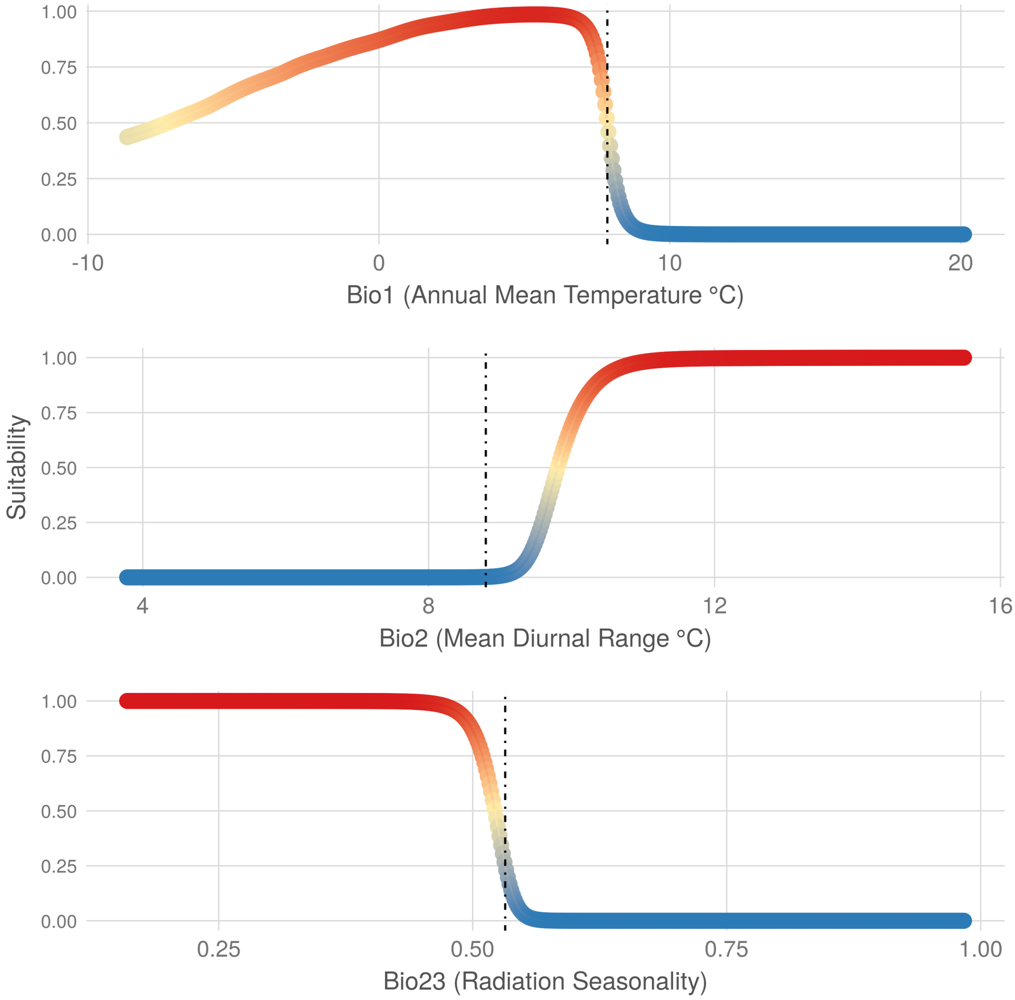


Figure 4: In this figure the average probability of presence (suitability) of *Abies alba* is plotted against the three variables with the highest average coefficient effect size in the model (top: range of annual mean temperature Bio1, middle: mean diurnal range Bio2, bottom: Radiation Seasonality or Bio23). The relationship between the probability of presence (suitability) of *Abies alba* and annual mean temperature has a “bell shape”, rising slowly moving from the left of the study area average (7.8 °C), peaking just before the average and decreasing rapidly when on its right. The shape of the relationship between the probability of presence and the mean diurnal temperature range is inverted. A low diurnal temperature range is associated with a low suitability while a wide temperature variability is associated with high suitability. The highest suitability is reported for Bio2 values higher than 11 °C. The Radiation Seasonality (the standard deviation of the weekly solar radiation estimates expressed as a percentage of the mean of those estimates) shows a negative pattern with respect to suitability. Areas with a very high average difference in solar radiation during the year (i.e. Northern Europe) are reported as weakly suitable for *Abies alba*. All the curves were obtained varying the value and the model coefficient of Bio1, Bio2 and Bio23 while keeping the values of the other predictors at their average. As reported in the main text, this results as well as that in Figure 5 is derived from the model with a strong prior on sampling effort.

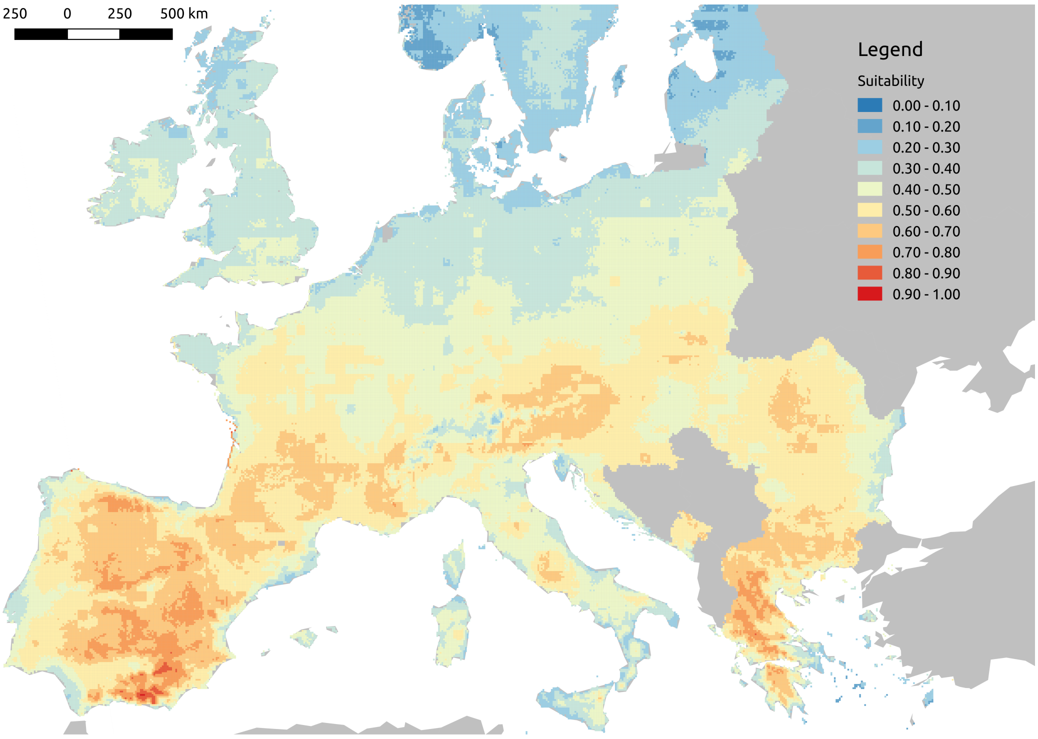


Figure 5: *Abies alba* suitability distribution as derived from the multi-level model with strong prior on sampling effort. The pixel value is the average of the PPDs for that pixel.

Figure 1

Click here to download high resolution image

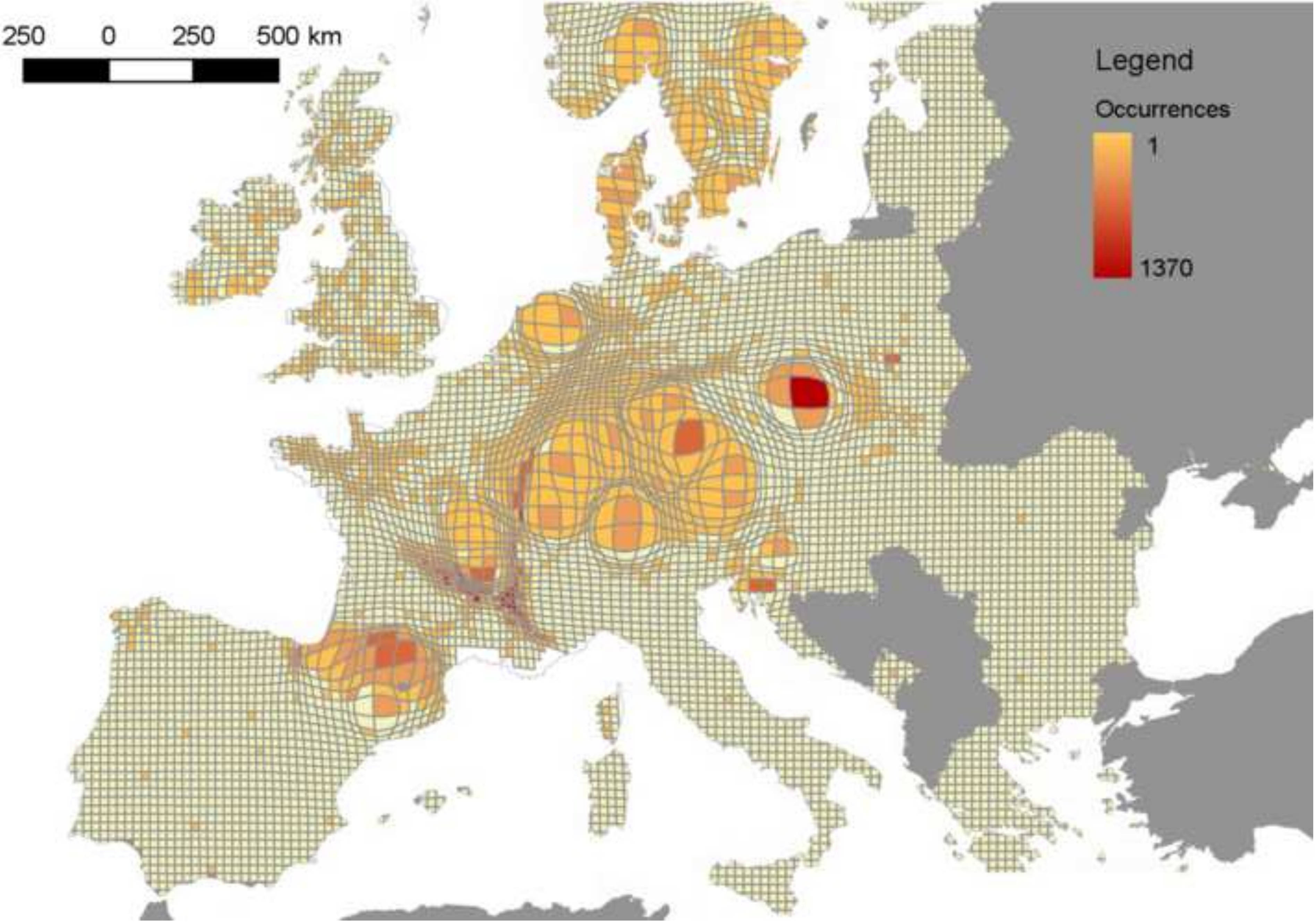


Figure 2

Click here to download high resolution image

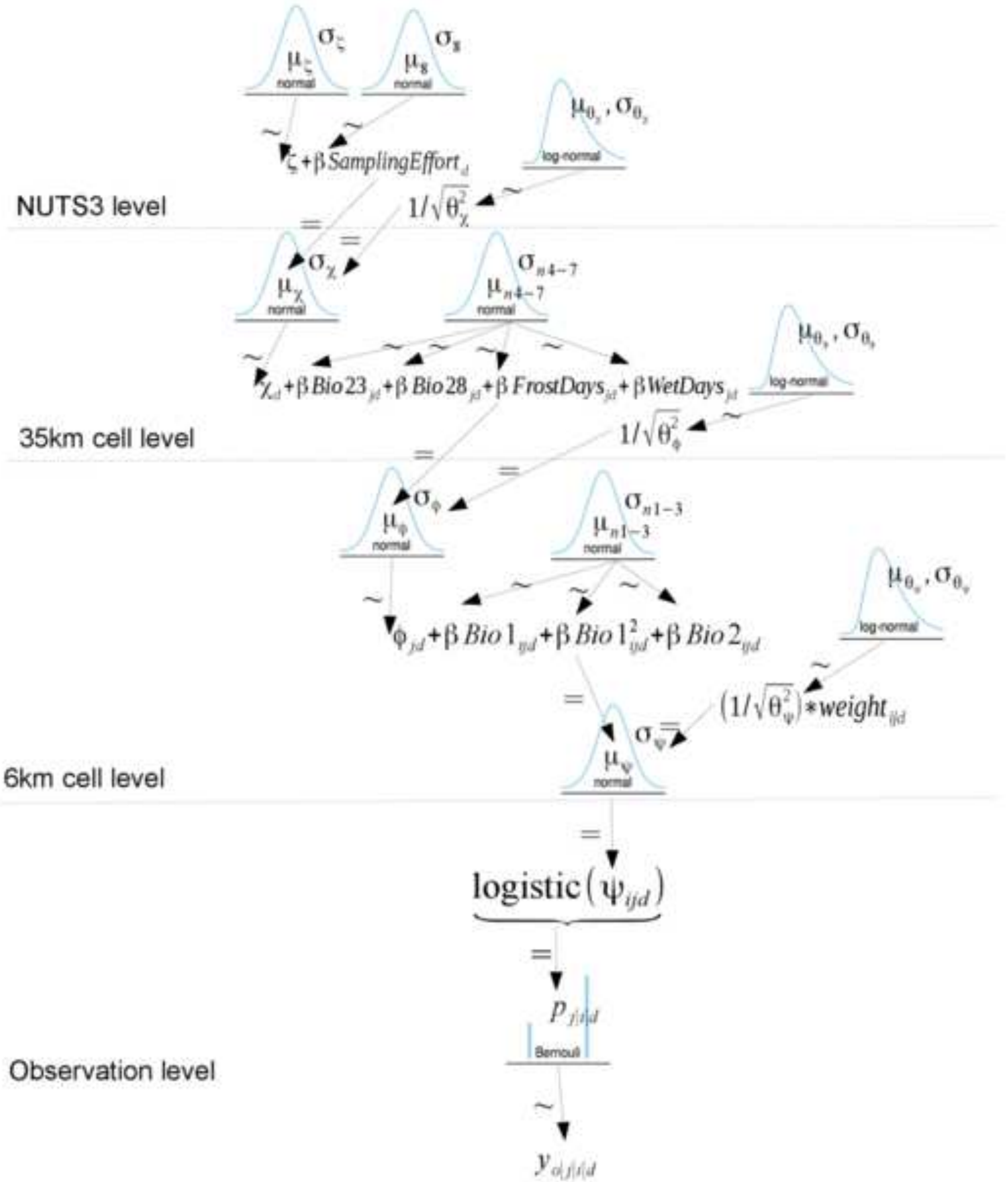

σ_ζ
μ_ζ
normal
σ_g
μ_g
normal
$\mu_{\theta_\chi}, \sigma_{\theta_\chi}$
log-normal
~
~
$\zeta + \beta\, SamplingEffort_d$
~
NUTS3 level
$1/\sqrt{\theta^2_\chi}$
=
=
σ_χ
μ_χ
normal
σ_{n4-7}
μ_{n4-7}
normal
~
~
~
~
~
$\mu_{\theta_\phi}, \sigma_{\theta_\phi}$
log-normal
$\chi_d + \beta\, Bio23_{jd} + \beta\, Bio28_{jd} + \beta\, FrostDays_{jd} + \beta\, WetDays_{jd}$
~
35km cell level
$1/\sqrt{\theta^2_\phi}$
=
=
σ_ϕ
μ_ϕ
normal
σ_{n1-3}
μ_{n1-3}
normal
~
~
~
~
$\mu_{\theta_\psi}, \sigma_{\theta_\psi}$
log-normal
$\phi_{jd} + \beta\, Bio1_{ijd} + \beta\, Bio1^2_{ijd} + \beta\, Bio2_{ijd}$
~
=
$(1/\sqrt{\theta^2_\psi}) * weight_{ijd}$
=
σ_ψ
μ_ψ
normal
6km cell level
=
$\underbrace{\text{logistic}(\psi_{ijd})}$
=
$p_{j|i|d}$
Bernoulli
Observation level
~
$y_{o|j|i|d}$

Figure 3

Click here to download high resolution image

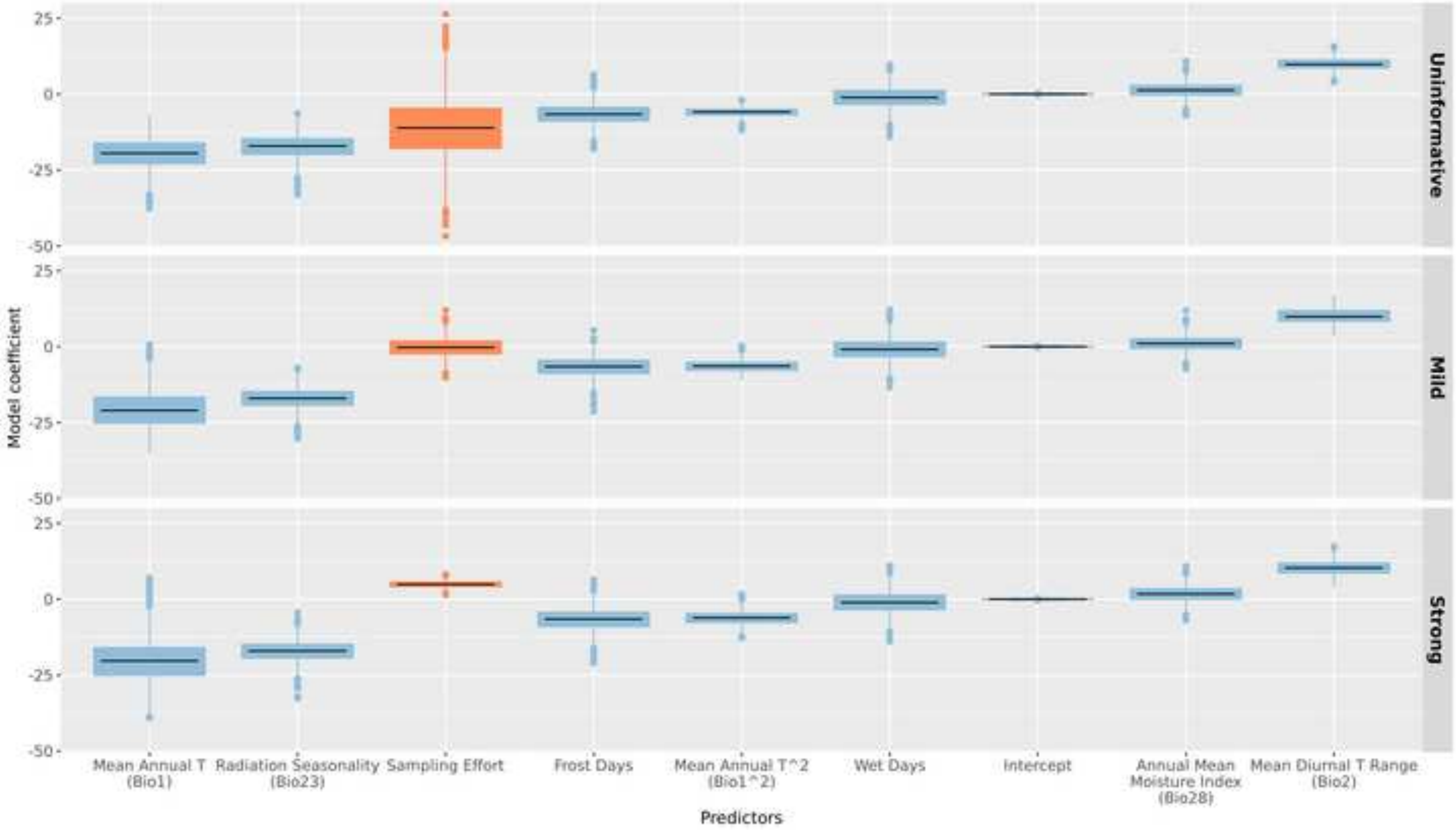


Figure 4

Click here to download high resolution image

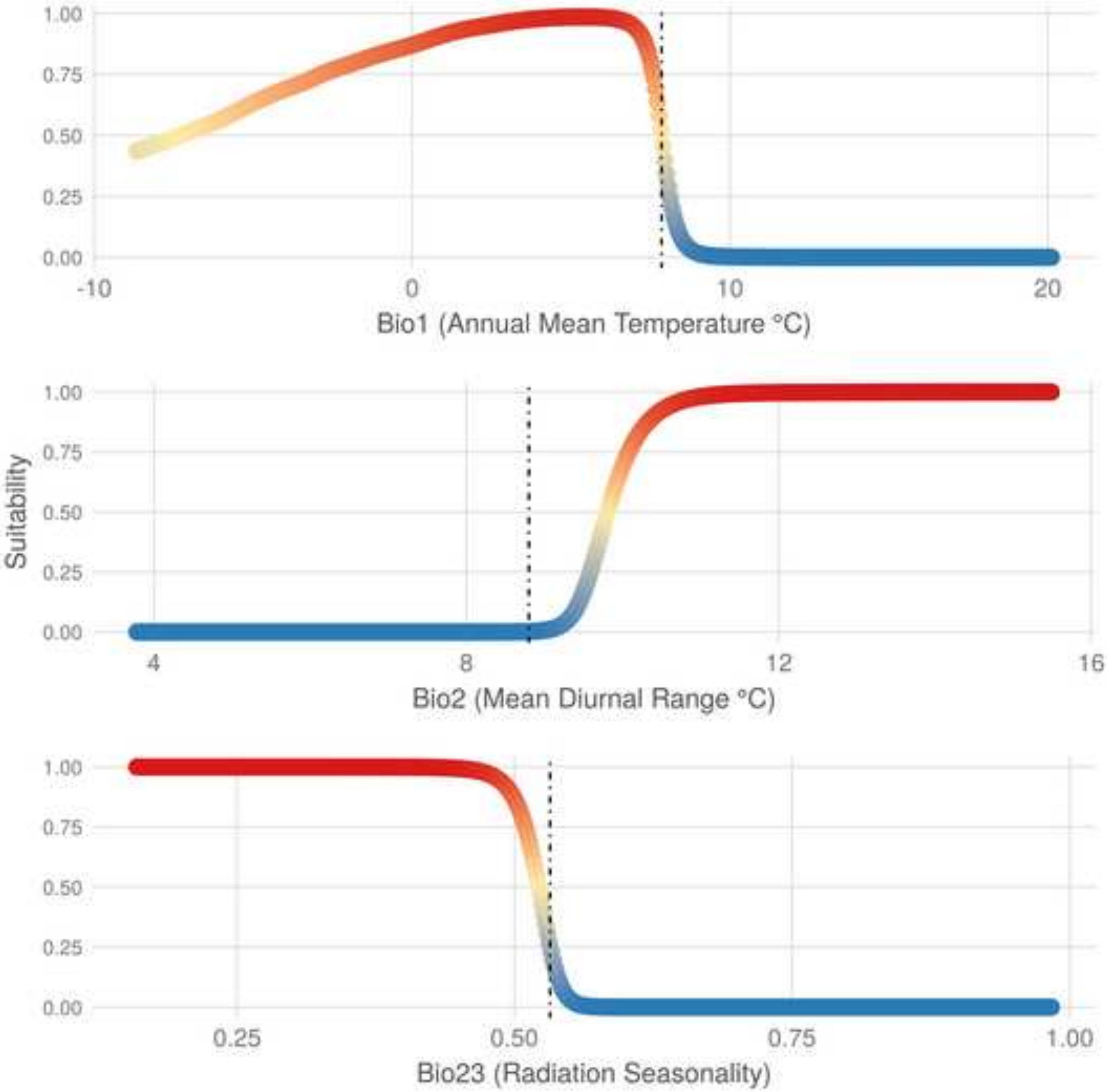


Figure 5

Click here to download high resolution image

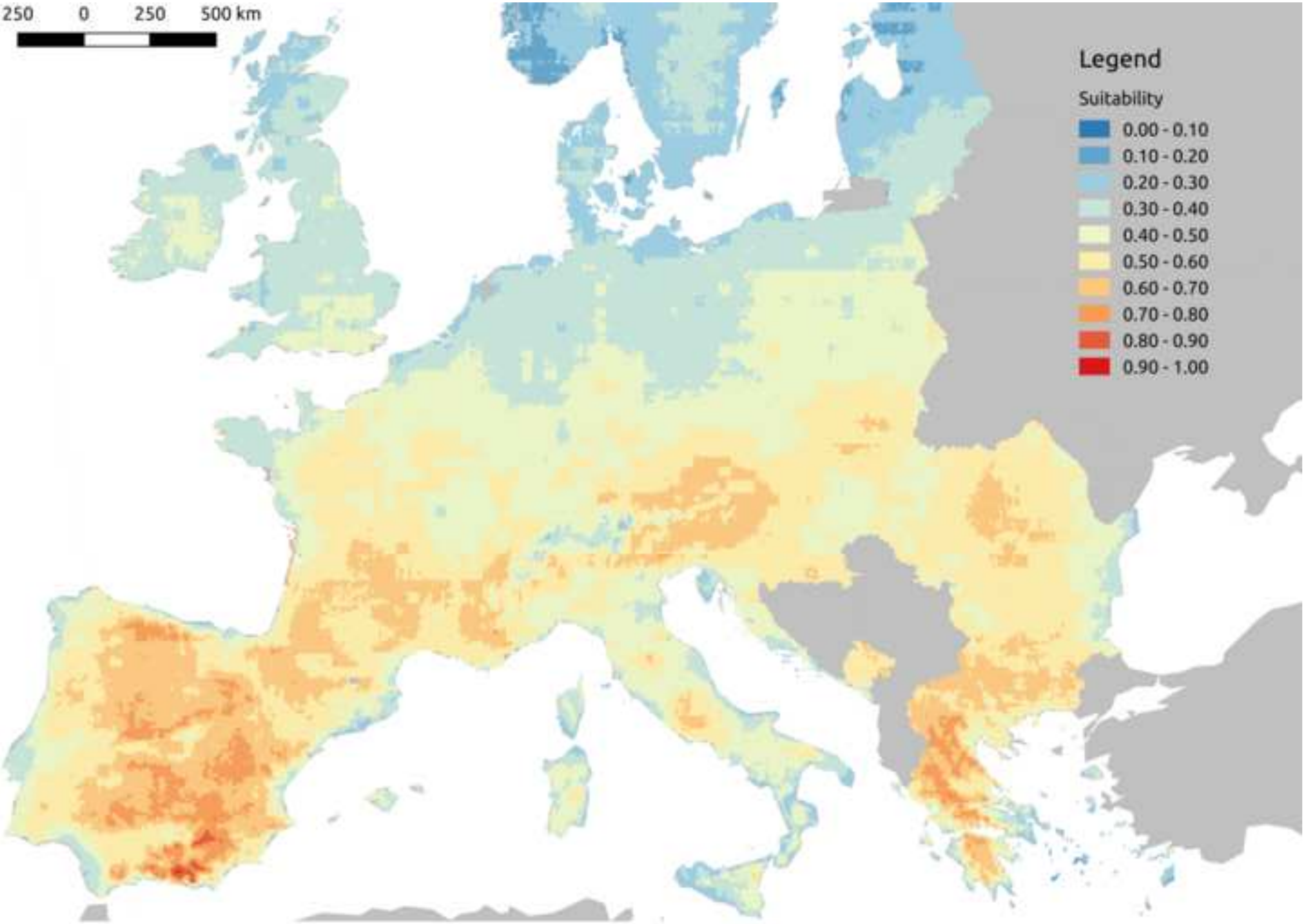


Figure S1

Click here to download high resolution image

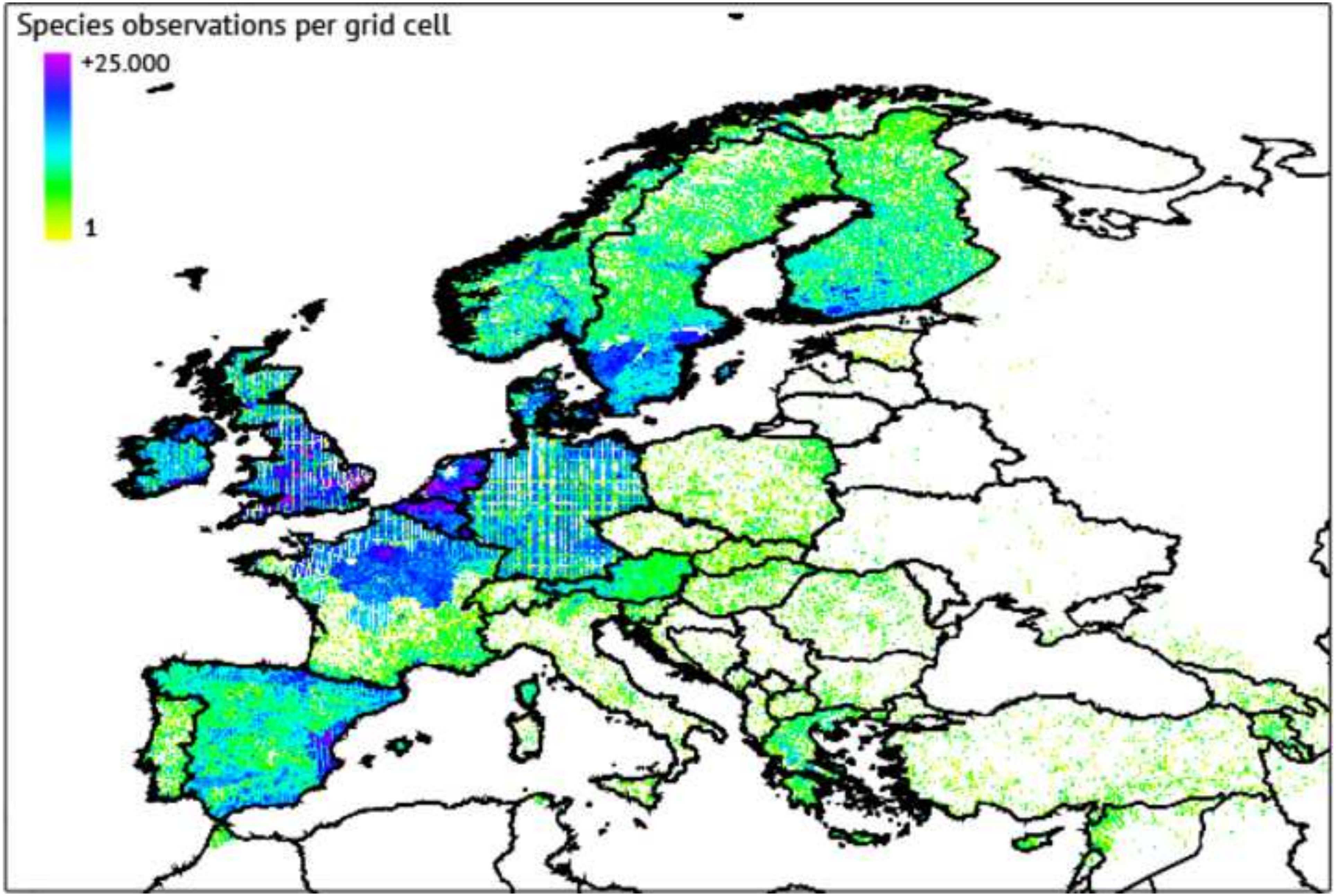


Figure S2

Click here to download high resolution image

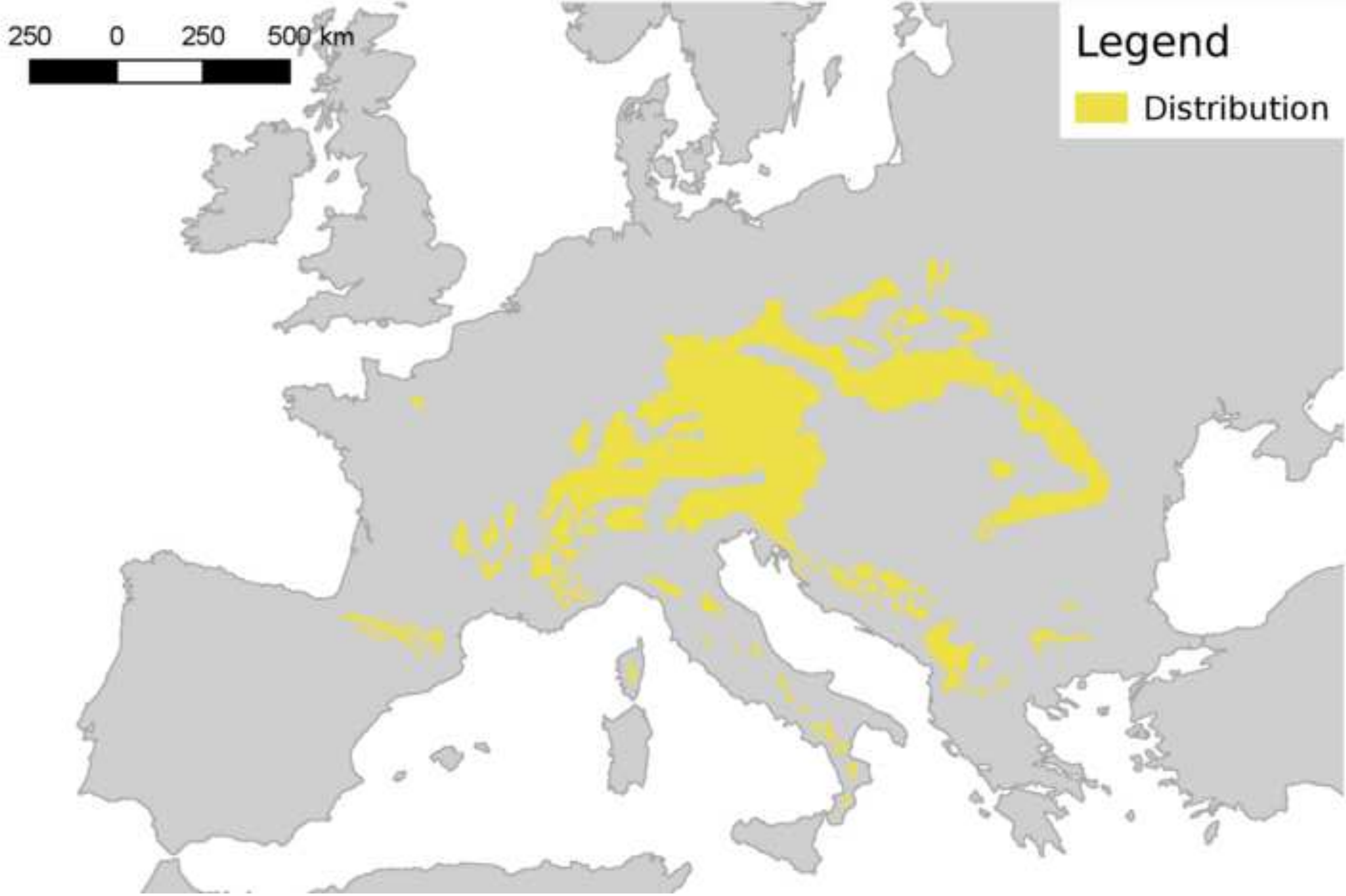


Figure S3

Click here to download high resolution image

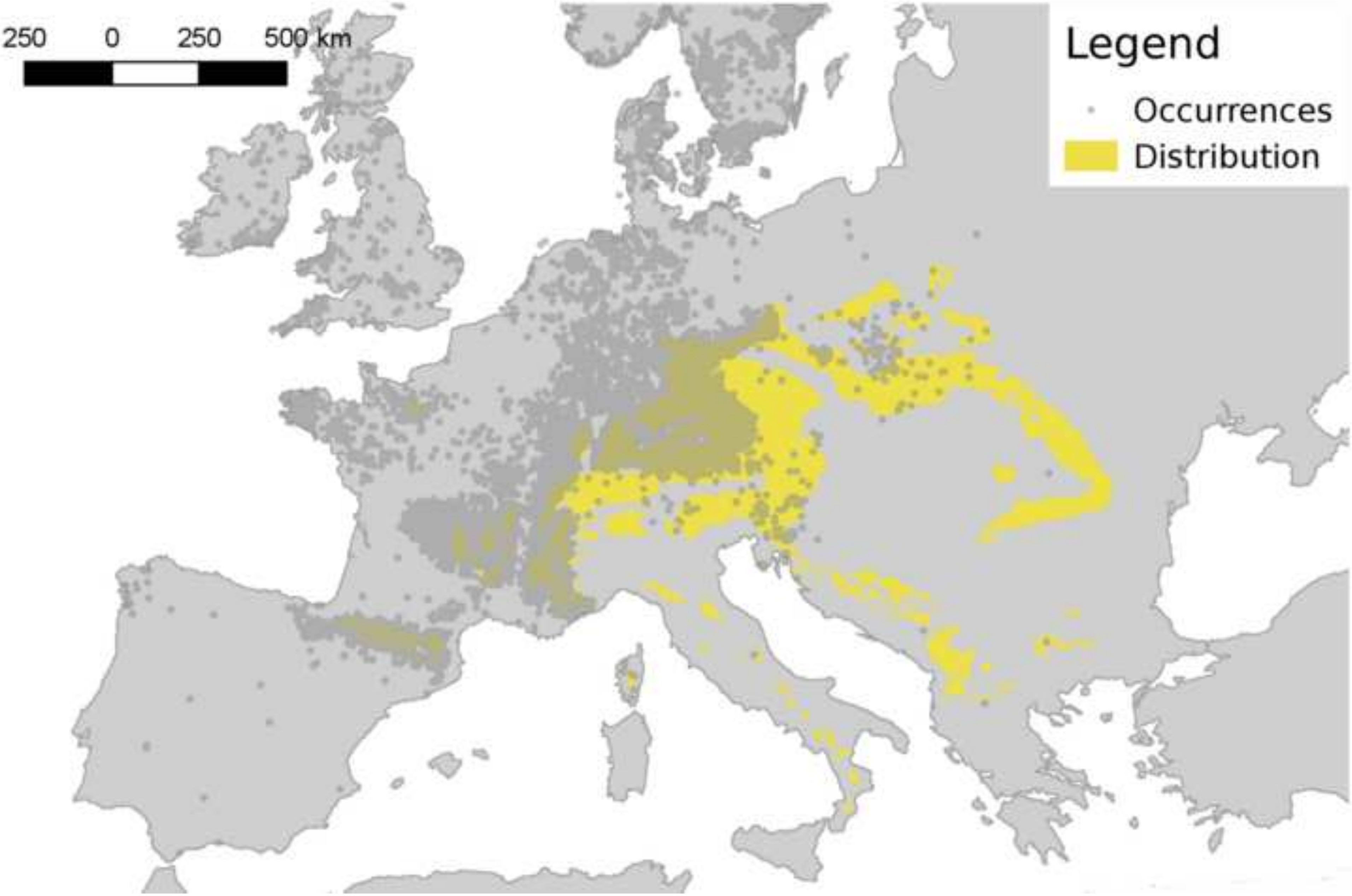
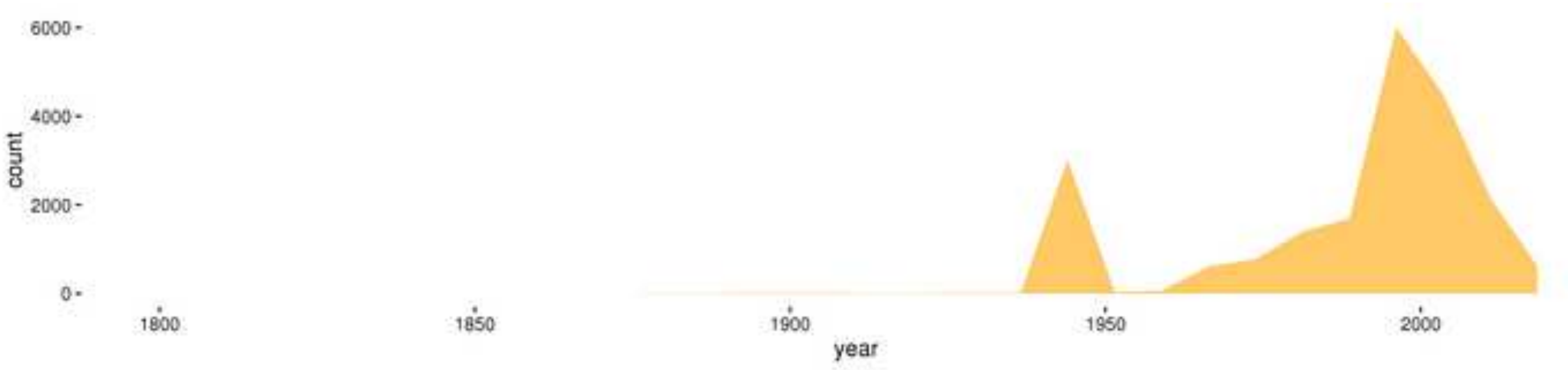


Figure S4

Click here to download high resolution image


count
year