Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2014) **23**, 457–467

A Journal of Macroecology

Global Ecology and Biogeography

RESEARCH PAPER

Where are the wild things? Why we need better data on species distribution

Anne Dupuié[1]*, Niklaus E. Zimmermann[2] and Isabelle Chuine[1]

[1]*CEFE UMR 5175, 1919 Route de Mende, 34293 Montpellier Cedex 5, France,* [2]*Landscape Dynamics Unit, Swiss Federal Research Institute, WSL, Zuercherstrasse 111, CH-8903 Birmensdorf, Switzerland*

*Correspondence: Anne Duputié, Laboratoire de Génétique et Évolution des Populations Végétales, UMR 8198, CNRS – Université Lille 1, Sciences et Technologies, Cité Scientifique, 59650 Villeneuve d'Ascq, France.
E-mail: anne.duputie@ens-lyon.org

ABSTRACT

Aim The effects of ongoing global change are causing increasing concern about the ability of species or biomes to shift or adapt. Tremendous efforts have been made to develop ever more sophisticated species distribution models to provide forecasts for the future of biodiversity. All these models rely on species occurrence data, either for calibration or validation. Here we evaluate (i) whether distribution data diverge among widely used sources, for supposedly well-known taxa, and (ii) to what extent these divergences affect species distribution models.

Location Europe (as an example).

Methods We compared the distribution maps of 21 of the most common European trees, according to four large-scale, putatively reliable sources of distribution data. For each species, we compared the outputs of correlative species distribution models built using occurrence data from each of these sources of data. We also investigated how discrepancies in large-scale occurrence data affected the validation scores of two process-based tree distribution models.

Results Maps of tree occurrence diverged in 8–74% of the forested area, depending on species. These discrepancies affected projections of niche models: for example, 22–75% of the area projected as suitable by at least one model generated using one source of data was not projected as such by all other models. For most species, this proportion increased under scenarios of climate change, whatever the model used. To a lesser extent, uncertainties on current species distributions also affect the validation score of process-based distribution models.

Main conclusions Reliable, widely used sources of occurrence data strongly diverge even for well-known taxa – the most common European trees. Scientists and stakeholders should acknowledge this gap in knowledge, since accurate data are a prerequisite to providing stakeholders with robust forecasts on biodiversity. Participatory science programmes and remote sensing techniques are promising tools for rapidly gathering such data.

Keywords

Climate change, European forest trees, model robustness, species distribution models, species occurrence data, uncertainty.

INTRODUCTION

Together with other components of global change, such as habitat fragmentation and land-use change, climate change contributes to the demographic decrease, the contraction and/or the shift of the geographic ranges of many species (Pereira *et al.*, 2010). Species distribution models (SDMs) are widely used to project the changes in species or biome distributions under changing environments. For example, the expected changes in the distribution of biodiversity (Morin *et al.*, 2008; Thuiller *et al.*, 2011) and the fate of nature reserves (Araújo *et al.*, 2011; Hickler *et al.*, 2012) under climate change have garnered much attention. Models generally agree that the suitable habitats for most species will shift generally, though not only, towards the poles (VanDerWal *et al.*, 2013), but the capacity of individual species to colonize new favourable areas will depend partly

DOI: 10.1111/geb.12118
http://wileyonlinelibrary.com/journal/geb

on their competitive and dispersal abilities (Boulangeat *et al.*, 2012; Zhu *et al.*, 2012).

SDMs can broadly be classified into two categories, although a continuum exists between the two (Dormann *et al.*, 2012): correlative SDMs establish statistical relationships between environmental variables and observed species occurrences and process-based SDMs describe the empirical reaction norms of physiological processes to environmental drivers to estimate species performance and presence. Models of the latter class do not normally use directly observed occurrences for calibration (forward process-based models; but see Higgins *et al.*, 2012) but do use such data for evaluation. In most cases, models of both types only inform about the suitability of a particular environment, regardless of extrinsic factors that may prevent a species from actually colonizing a site, such as competitive exclusion or dispersal limitation (Dormann *et al.*, 2012). In addition, models driven by distribution data (i.e. correlative SDMs) inform on correlations between environmental variables and species distributions, which may be causal but may also be contingent to the past history of the species and its environment (e.g. Hortal *et al.*, 2012).

Projections of models of both types are affected by various intrinsic and extrinsic factors. Among these, the incompleteness of models (Dormann, 2007), the extent of environmental differences between the novel and the calibration environments (Veloz *et al.*, 2012) and the uncertainties in scenarios of land use and climate change (Buisson *et al.*, 2010) clearly matter. Correlative SDMs suffer from additional sampling bias of the occurrence (and absence) data used for calibration, such as sample size (Stockwell & Peterson, 2002), spatial scale (Randin *et al.*, 2009), location error (Guisan *et al.*, 2007), species detectability (Reese *et al.*, 2005), biases in sampling effort (Lobo, 2008) and the relative extent of the modelled area actually occupied by the species (Jiménez-Valverde *et al.*, 2008). Finally, the validation of any SDM has a deep reliance on the quality of the 'reference' presence and absence data.

If issues of data quality for the calibration of correlative SDMs have been widely addressed in the literature, no study has so far evaluated the impact of the potential discrepancy among several data sources on the projections of SDMs. Divergences among sources of occurrence data may be thought to affect primarily species that are rare, inconspicuous or both. Here we demonstrate dramatic differences in data obtained from different, putatively reliable sources on the occurrence of widespread and conspicuous species: the most common European forest trees. We explore to what extent these discrepancies affect the outputs of correlative (driven by distribution data) SDMs, and the validation of process-based SDMs.

MATERIAL AND METHODS

Sources of occurrence data

We used four sources of tree distribution data covering the whole European continent: the Atlas Florae Europaeae (AFE; Jalas & Suominen, 1964–2010), the map of the natural vegetation of Europe (EuroVegMap; Bohn *et al.*, 2004), the EUFORGEN database (http://www.euforgen.org/distribution_maps.html) and the Joint Research Centre distribution maps (JRC; http://forest.jrc.ec.europa.eu/). In addition, punctual occurrence data were obtained from forest inventory plots (6146 plots from the ICP level I dataset; http://www.icp-forests.org/). These were used to evaluate the accuracy of the atlas-based data.

The JRC dataset provides densities of tree species within European forests only, at 1 km resolution. This dataset is partly based on the ICP forest plot inventory, which it matched perfectly. However, because the JRC dataset only covers forested areas, it is expected to underpredict species occurrences in non-forest areas (where trees can grow outside forests, or in small forest patches).

Atlas Florae Europaeae is a project launched in 1965 and still ongoing. It is based on the field work of botanists across Europe. This atlas means to exhaustively cover the European flora, at a coarse resolution (50 km × 50 km pixels). This source of data was thus expected to overpredict species presence, especially in contrasted areas (such as mountain ranges). As compared with the ICP forest inventory data, AFE indeed showed a large number of false positives (i.e. it overpredicted distributions of many species; see the first column of the table in Appendix S1 in the Supporting Information).

EuroVegMap is a project launched in 1975 and completed in 2004. Its aim was to draw a fine-scale map of the potential vegetation of Europe. As such, the mapping project focused on habitats rather than species. A number of habitats were defined, each of which was characterized by the presence of some species. Not all species present in a given habitat are present in all of its patches. As such, we expected this map to overpredict the distribution of at least some species. Yet, surprisingly, the rate of false positives (overpredicted occurrences, as compared with the forest inventory data) was often less than for the other two atlases (Appendix S1).

The EUFORGEN database was created by the European Forest Genetic Resources Programme. Distribution maps are based on both bibliographic work and expert knowledge. We did not have any particular expectations as to the rates of over- or underprediction of this source of data. For many species, this large-scale dataset is the one most able to correctly predict occurrences, as indicated by the ICP forest inventory (Appendix S1), but this may be because ICP data were partly used to compile this dataset (no information is available as to the sources of this dataset).

Because they are based on national or international collaborations of experts and aim at extensive coverage, such sources of data are expected to suffer little sampling bias. They are thus often used as input for correlative SDMs (e.g. Araújo *et al.*, 2011; Thuiller *et al.*, 2011), and are also used to validate distribution models (e.g. Hickler *et al.*, 2012; Gritti *et al.*, 2013).

Species distribution data

Distribution data were collected from these five sources of data for 21 of the most common European forest tree species: *Abies*

alba Mill. (silver fir), *Acer campestre* L and *Acer pseudoplatanus* L. (field and sycamore maples), *Alnus glutinosa* (L.) Gaertn. (black alder), *Betula pendula* Roth and *Betula pubescens* Ehrh. (silver and white birches), *Carpinus betulus* L. (common hornbeam), *Castanea sativa* Mill. (chestnut), *Corylus avellana* L. (common hazel), *Fagus sylvatica* L. (European beech), *Fraxinus excelsior* L. (common ash), *Larix decidua* Mill. (European larch), *Picea abies* (L.) H. Karst. (European spruce), *Pinus halepensis* Mill., *Pinus nigra* J.F.Arnold, *Pinus pinaster* Aiton and *Pinus sylvestris* L. (Aleppo, black, maritime and Scots pines, respectively), *Populus tremula* L. (quaking aspen), *Quercus ilex* L., *Quercus pubescens* Willd. and *Quercus robur* L. (holm, downy and pedunculate oaks, respectively). Whenever a species was split into subspecies, occurrences of all subspecies were merged to yield the final occurrence datasets.

Data from AFE were not available for three species (*Acer campestre*, *Acer pseudoplatanus*, *Fraxinus excelsior*); data from EUFORGEN were not available for five species (*Betula pubescens*, *Carpinus betulus*, *Corylus avellana*, *Quercus ilex* and *Quercus pubescens*), and data from JRC were not available for three species (*Acer campestre*, *Acer pseudoplatanus*, *Corylus avellana*). Each species was thus represented by two (*Acer campestre* and *Acer pseudoplatanus*) to four (13 species) maps.

Distribution data obtained from the five sources of data were either upscaled or downscaled to the resolution of 10′ (see Appendix S2). AFE occurrence data (whether native or alien) were downscaled to 10′ by attributing occurrences from one 50-km AFE cell to all 10′ pixels overlapping this cell. This necessarily leads to overestimations of species ranges (see Appendix S3; also discussed in Rondinini *et al.*, 2006; Pineda & Lobo, 2012). The EUFORGEN dataset consists of continuous areas and punctual occurrences: occurrences were attributed to each pixel of the 10′ grid overlapping a continuous area, or containing at least one punctual occurrence. The EuroVegMap dataset is provided as a series of polygons corresponding to potential vegetation types (including plantations) at a resolution of 2 km. Species appear as 'present' in all patches of the types of habitats they are known to inhabit, hence over whole polygons of habitats. For each species, we attributed 'occurrence' records to each of the 10′ pixels totally or partly overlapping one such unit. JRC data record species abundances within forests, at 1-km resolution. JRC data were transformed into presence–absence data by attributing an 'occurrence' record to each 10′ pixel overlapping at least one 1-km JRC pixel with positive abundance. Out of the 28,766 $10' \times 10'$ pixels used in this study, 10,296 contained forest patches and thus information from the JRC dataset. The ICP dataset provides punctual presence/absence data in 6146 plots; presences or absences were attributed to the 10′ pixel overlaying the location of the initial record. Whenever several ICP plots were included in the same pixel, presence was attributed to the $10' \times 10'$ pixel as soon as the species was present in at least one of the ICP plots. This resulted in 5441 $10' \times 10'$ pixels containing ICP information. The upscaling and downscaling procedure for data from all sources is illustrated in Appendix S2.

Climate data

Five climatic variables were used to build correlative distribution models. These variables are related to temperature and precipitation, and their seasonality. They thus reflect constraints on energy and water uptake, and on temperature and water stress, which are thought to play an important role in limiting species distributions. These variables were: minimal mean temperature of the coldest month (°C); mean yearly growing degree days above 5 °C, a rough indicator of potential plant energy uptake (°C); mean total yearly amount of precipitation (mm); seasonality of precipitation, as expressed by the coefficient of variation of precipitation across the four trimesters (dimensionless); and a dimensionless moisture index, expressed as the ratio of mean actual to potential evapotranspiration over the growing season (temperatures above 5 °C).

Current (1980–2000) climate variables were computed from the CRU TS 1.2 dataset and forecasts (2080–2100) were derived from the CRU TYN SC 1.0 dataset. To isolate the impact of data-driven uncertainty on SDMs, we deliberately used a unique general circulation model (HadCM3). We considered two SRES emissions scenarios: A1Fi and B2. Scenario A1Fi implies more emissions (and thus higher concentrations of atmospheric CO_2) than scenario B2.

Correlative distribution modelling

Correlative habitat suitability models were generated for each species and each available source of data, with the current climate dataset. Because some of the sources of data used here were atlases mapping regions of occurrence, but not providing punctual occurrences, models relying on the density of presences only (such as MaxEnt or Poisson regression models) would necessarily have been imprecise. For this reason, we used presence–absence models, implemented in the Biomod library in R (Thuiller *et al.*, 2009). For each species and source of data, five algorithms were run: artificial neural networks (ANNs), classification tree analysis (CTA), flexible discriminant analysis (FDA), generalized additive models (GAMs) and generalized linear models (GLM). MaxEnt was not used for the reason given above; in addition, using it would have been somewhat equivalent to using a GLM (Renner & Warton, 2013).

These algorithms require occurrences and absences to be specified. For each species and source of occurrence data, we considered that the species may be absent wherever it was not observed, and generated pseudo-absences in locations where the species was not deemed as 'present'. This approach has been proved fruitful for virtual species (Wisz & Guisan, 2009; Barbet-Massin *et al.*, 2012). We only generated as many pseudo-absences as there were 'presence' records (Barbet-Massin *et al.*, 2012), and hence did not assume that all non-presences were true absences. This procedure was repeated three times for each species and source of data. Models were then calibrated using a random set of 70% of the available data (presences and pseudo-absences), and evaluated against the remaining 30% of the dataset, using the area under the receiver operating curve (AUC)

criterion (Swets, 1988). The AUC varies from 0 to 1, with 0.5 indicating a null model and 1 a perfect model. This criterion merely provides an evaluation of the model's discriminatory power (Lobo *et al.*, 2008). This procedure was repeated three times, to provide three-fold internal cross-validation. Cross-validation scores were generally high, regardless of the choice of pseudo-absences and the data splitting.

Habitat favourability is monotonously related to the continuous output of each algorithm, hereafter referred to 'habitat suitability' (Real *et al.*, 2006, equation 7). For each combination of species, source of data and model (algorithm), we determined a threshold above which the model was considered to project the species as 'present'. We chose this threshold so as to maximize the sum of sensitivity and specificity (Jiménez-Valverde & Lobo, 2007). We then weighted binary projections of the five models according to each model's performance and averaged them to yield an 'ensemble model'. We finally applied a threshold to the probability of occurrence generated by the ensemble model, again maximizing the sum of sensitivity and specificity, to compute statistics such as range size. Because we had access to neither true absence data nor to the geographic variation in sampling effort, at this stage we had to assume that non-presence points were absences.

Once calibrated on current climatic conditions, models were extrapolated to forecast conditions. For each species, source of data and scenario, we defined the area of the modelled species range as the cumulated surface (in km^2) of all pixels for which the ensemble model produced outputs were above that threshold. Range change under scenarios was computed as the ratio of forecast to current modelled area. We considered two extreme scenarios: no dispersal and full dispersal (Appendix S3).

The (exact) variance between projections was computed using the continuous outputs of the ensemble models generated using different data sources. When three or four data sources were available, variance could not exceed 1/3; when there were only two data sources (both *Acer* species), variance could reach 1/2. Inter-projection variance was summed over species after dividing variances for *Acer* species by 3/2, in order to obtain equal weights for all species.

Process-based distribution modelling

For three species (Scots pine, pedunculate oak and European beech), we ran a process-based distribution model (PHENOFIT; Chuine & Beaubien, 2001; Morin & Chuine, 2005), and a hybrid distribution model (LPJ; Smith *et al.*, 2001; Sitch *et al.*, 2003). Details about the models and their parameterization are provided in Gritti *et al.* (2013).

These models are not calibrated using distribution data; however, observed distributions are usually used to validate them. The quality of the projections obtained with these models was assessed using the AUC (Swets, 1988) criterion, using each source of data as a reference. For each of the process-based models, species, source of occurrence data and scenario, we determined the suitable area. 'Suitable' pixels corresponded to those where the model output was superior to a threshold maximizing the sum of sensitivity and specificity, with respect to the source of occurrence data (Jiménez-Valverde & Lobo, 2007).

RESULTS

Discrepancies among sources of occurrence data

We found surprisingly large discrepancies among the four sources of occurrence data (Appendices S4 & S5). Depending on the species, occurrence maps diverged for 8–74% (median 23%) of the forested area (only forests were exhaustively covered by all four data sources).

These differences translated into discrepancies in spatial patterns of species richness (Fig. 1) and affected most species across the whole continent (Appendices S4 & S5). Among the three sources of occurrence data covering the whole of Europe (AFE,

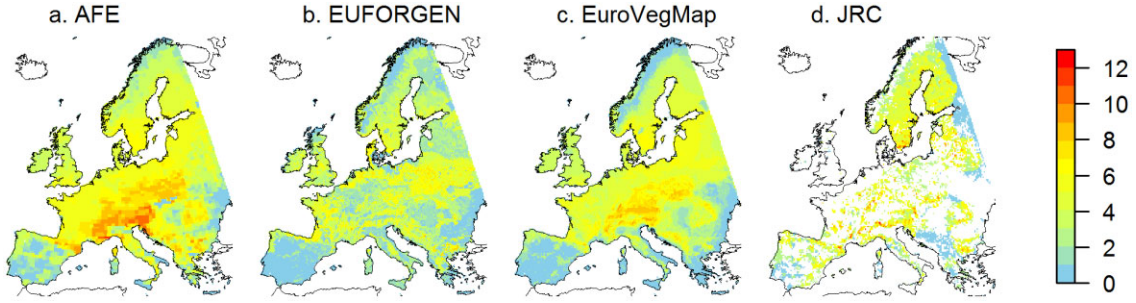

Figure 1 Spatial distribution of species richness (species count) for 13 common European trees, according to four sources of data (Lambert azimuthal equal area projection). Darker shades indicate higher richness. The 13 species included here are those of the 21 most common European species for which all four maps were available. Note that the JRC dataset (d) only covers forested areas (non-forest areas appear in white). AFE, Atlas Florae Europaeae (Jalas & Suominen, 1964–2010; Lahti & Lampinen, 1999; http://www.luomus.fi/english/botany/afe/index.htm); EUFORGEN, Euforgen dataset (http://www.euforgen.org/distribution_maps.html); EuroVegMap, map of the potential vegetation of Europe (Bohn *et al.*, 2004); JRC, Joint Research Centre dataset (http://forest.jrc.ec.europa.eu/).

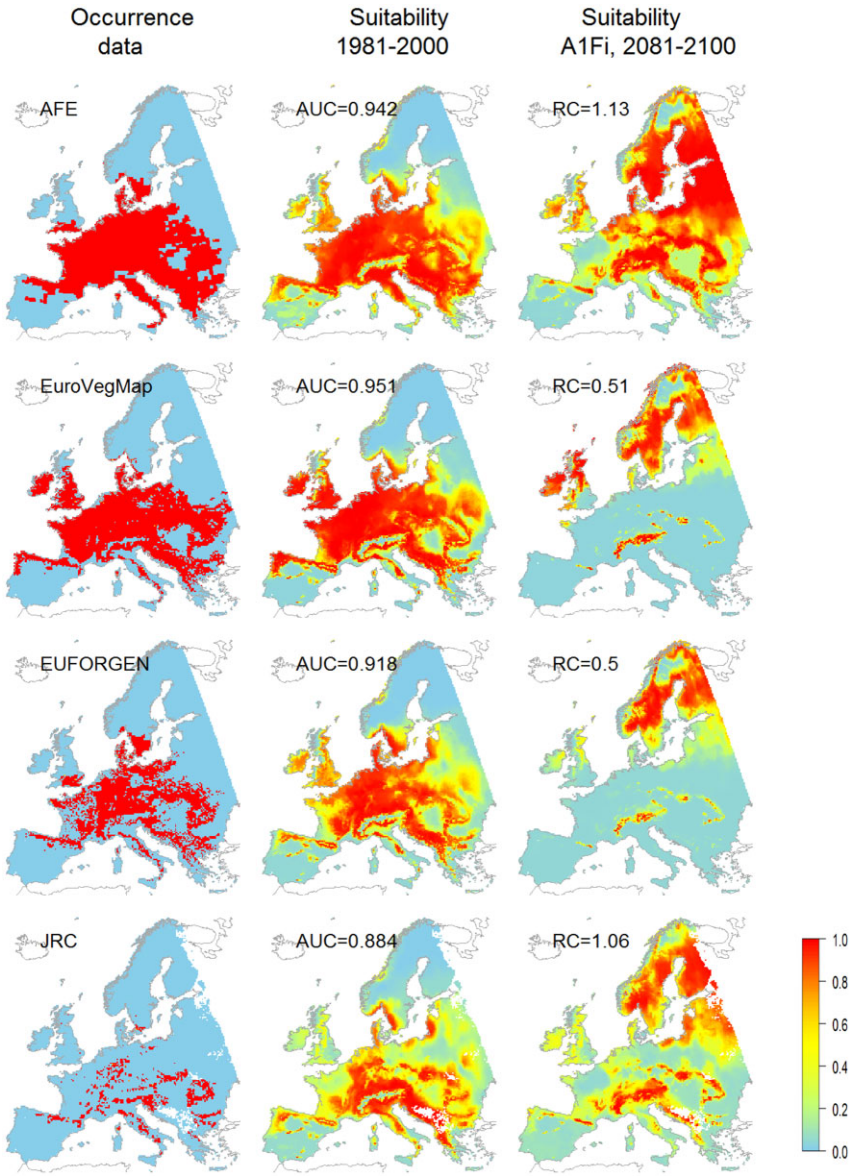

Figure 2 Consequences of discrepancies in sources of occurrence data on projections of correlative distribution models: the case of European beech, *Fagus sylvatica* (Lambert azimuthal equal area projection). Differences in occurrence data from four sources (left column, red pixels) translate into differences in current (middle column) and forecast (right column) simulated habitat suitability. AUC (the area under the receiver operating curve) indicates the discriminative power of each model (as compared with the corresponding occurrence map); values above 0.9 are considered as very good agreement. RC, range change, i.e. the ratio of future to current simulated suitable areas (assuming full dispersal).

EuroVegMap and EUFORGEN), the one with the coarsest resolution (AFE) consistently indicated wider distributions than the other two (Appendices S1 & S3); however, none of them was able to consistently better predict punctual forest inventory data (ICP dataset; Appendix S1).

Impact on projections of correlative SDMs

Whatever the species, source of distribution data or modelling algorithm, modelled distributions of current habitat suitability closely matched the source of distribution data, resulting in large discrepancies in simulated current habitat suitability. As an example, Fig. 2 shows how outputs of the BIOMOD model are affected by the source of data for the emblematic European beech (*Fagus sylvatica* L.; see Appendix S6 for all 21 species).

When extrapolated to climatic scenarios, models generated using different sources of occurrence data for a given species generally agreed in the direction of range shift, but showed large quantitative variance. For example, consistent with previous studies (Kramer *et al.*, 2010; Cheaib *et al.*, 2012; Meier *et al.*, 2012), all models generated using BIOMOD inferred a north-easterly shift of the distribution of European beech by 2080–2100 under the A1Fi scenario; yet simulated suitable future habitats may cover either a larger or smaller range than the current range (Fig. 2). The same applied to 6 of the 20 other species under scenario A1Fi (and eight under scenario B2; Appendices S3 & S6). Thus, discrepancies among maps generated for current conditions were magnified under forecasts.

When using BIOMOD for the current period, depending on species, 22–75% (median 44%) of the area projected as 'suitable' by at least one model based on one source of occurrence data was not projected as such by the other models. This proportion increased for 19 (respectively 12) of the 21 species by 2080–2100 under scenario A1Fi (respectively B2; Appendix S7). Maps of

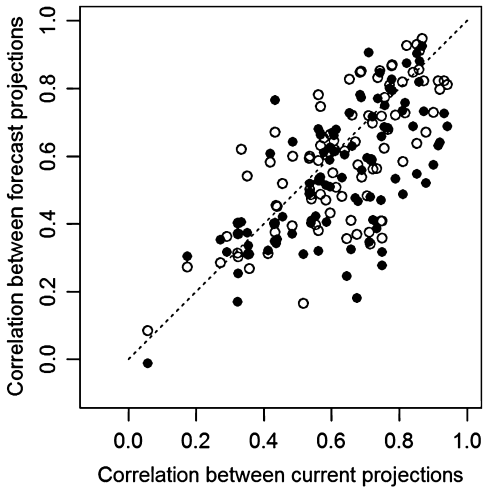

Figure 3 Discrepancies between model projections due to the source of occurrence data used for calibration increase under climate change scenarios. Species-wise correlations (r) between pairs of projected habitat suitability maps for 2080–2100 (closed symbols, A1Fi scenario; open symbols, B2 scenario) are plotted against the correlations between pairs of current habitat suitability maps obtained using different data sources as model input. On each panel, each point represents the correlation between two source maps, for a given species; and the dashed line is the 1 : 1 curve.

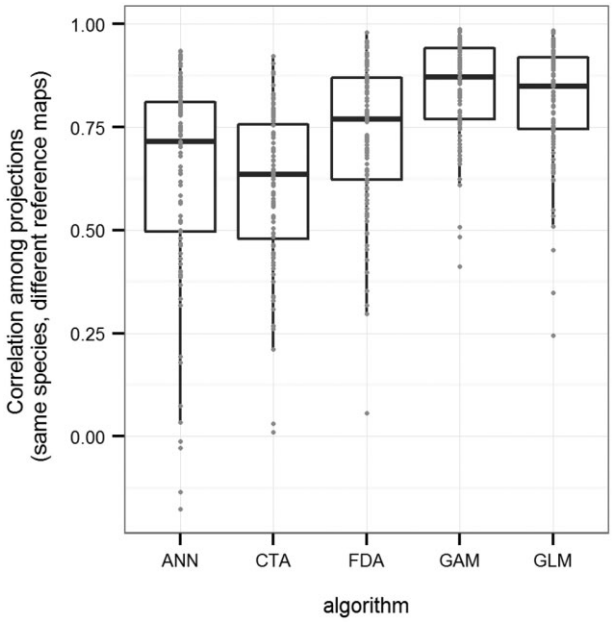

Figure 4 Distribution of the correlations among maps of habitat suitability generated for the period 1981–2000 by the five algorithms implemented in BIOMOD, for the same species but from different sources of occurrence data. Each dot represents the correlation between two projections made for a given species, but from different maps.

suitable habitats generated for the same species, but from different sources of data, showed overall low pairwise correlation under current conditions, and these correlations were even lower under forecast conditions (Fig. 3; e.g. the proportion of pairwise correlations r lower than 0.5 was 23% for current conditions and 39% for forecasts).

Among algorithms, GAMs and GLMs were the least sensitive to discrepancies among sources of occurrence data (Fig. 4).

Impact on the validation of projections of process-based SDMs

To a lesser extent, differences between sources of occurrence data also affected post-hoc quality assessments of process-based distribution models. The value of the AUC computed for three species and with two process-based models varied by 0.05–0.15 units depending on the data source chosen as a reference (Appendix S8), thus affecting judgment on the quality of these models. When outputs of process-based models were transformed into binary presence–absence projections, using a threshold maximizing the sum of sensitivity and specificity with respect to one reference source of data, the area projected as 'present' varied on average by 26% for the current period (Appendix S9).

DISCUSSION

SDMs need to be calibrated and/or validated against observed distribution data. Our results show that existing data in widely used sources show large discrepancies, even for well-known species in well-sampled areas such as forest trees in Europe. These discrepancies affect the projections of correlative distribution models (whether making use of presence-only or of presence/pseudo-absence data), especially under forecast climatic conditions, and also, to a lesser extent, the post-hoc validation score of process-based distribution models.

Why do sources of occurrence data differ?

Differences among sources of occurrence data may be attributed to various causes. First, the aims of the different mapping projects varied, thus the criteria used for mapping differed between projects. In particular, some datasets were assembled by botanists (AFE, EuroVegMap), others by foresters (EUFORGEN, ICP) and still others by foresters and modellers (JRC). Differences in interest may entail differences in mapping, and some investigators may consider the species as present only if it reaches some density, or if it is able to naturally regenerate. Second, all these global sources of data were gathered by many investigators located in different regions, and large-scale distribution maps are likely to suffer from regional bias due the subjectivity of local investigators. Finally, not all maps are provided with the same spatial resolution. All sources of data used here focus on presences, not on absences, hence absence data are often likely to be false absences. Datasets with a coarser grain were thus expected to overpredict presences – to some extent, this was true of the Atlas Florae Europaeae dataset.

Ideally, SDMs would need to be informed either by densities of observations (for models such as MaxEnt or Poisson regression models) or by absence data. True absence data are difficult to collect, even for easily detectable and immobile species such as trees, because they require a high sampling effort. Thus, large-scale sources of distribution data such as those used here only inform on occurrences, not on absences. Absence data are uncertain (Lobo *et al.*, 2010; Peterson *et al.*, 2011; Rocchini *et al.*, 2011). Some uncertainties pertain to the species' detectability, which in our case is supposedly large – we only selected common tree species on a continent long inhabited and heavily managed by humans. Some uncertainties pertain to grain size, which is often coarse in atlases (Rocchini *et al.*, 2011). Others are related to sampling biases or errors in location or in species identification. For example, point occurrence data (such as those provided by herbarium samples or the Global Biodiversity Information Facility) may be imprecise (Guisan *et al.*, 2007) or may have different densities resulting from differences in data collection or publishing effort (Kadmon *et al.*, 2004; Yesson *et al.*, 2007; Peterson *et al.*, 2011). Finally, large-scale atlases of distributions may be more precise or more exhaustive for some regions (and/or for some species) than for others. In well-sampled areas, absences may thus reflect the distribution of true absences, while in less-sampled areas, absences are more likely to be false absences. While it may be fruitful to inform a model by the sampling effort (Lobo, 2008), this information is not always available – and was not available in the present study. To be fully compatible with distribution modelling, large-scale maps of species distribution should thus be associated with maps of sampling effort.

To complicate matters, SDMs should only be calibrated using occurrences of viable populations. This is one of the reasons why GBIF data were not included in our analysis. GBIF records mix naturally regenerating populations with populations living in artificial areas, where they may never regenerate – or regenerate only owing to being watered or fertilized. A second important reason for excluding GBIF data from our analyses is that GBIF data also suffer from regional bias in sampling design or recording intensity (Yesson *et al.*, 2007). For example, for the species considered here, the UK, Netherlands and central/eastern parts of France were heavily sampled, while few occurrences of any tree species were recorded from Poland; and occurrence data for Spain were placed on a grid much coarser than the 10′ resolution used in this study. In their present state, GBIF data may be used to answer some questions for some taxa (e.g. Randin *et al.*, 2013), but this needs careful pre-filtering of the data (Beck *et al.*, 2013; Randin *et al.*, 2013).

Combining sources of occurrence data

Because none of the sources of data could arguably be considered as consistently better than the others in depicting observed occurrences (Appendix S1), none of the forecasts presented in Fig. 2 (nor in Appendix S6) can be deemed to be most plausible. Among the algorithms used, GAMs and GLMs seem to be the less sensitive to variations in source data. GLMs were also found to be relatively robust (as compared with other modelling algorithms) to small locational errors (Guisan *et al.* (2007). However, our study and that of Guisan *et al.* rely on a single climatic dataset; this result thus might not be general.

Combining the information conveyed by all data sources to obtain 'ensemble datasets' of distribution data that would approach the true empirical distribution of the species may prove tricky. For example, one could consider the species to be present only in sites where all data sources indicate presence; but this would lead to large numbers of false negatives (sites where the species would be falsely inferred as absent). The best combination of occurrence maps may be a map of the probability density of the given species being present. In our case, there is a 100% probability that species are present where the forest inventory dataset (ICP forests) indicates it to be. However, absences in this dataset may be false absences. In all pixels where the ICP forest dataset indicates no presence (or provides no data), the probability of the species being present may be defined as the probability of each of the other sources of data indicating it is present.

These kinds of 'ensemble data' could be used in SDMs. Indeed, these models could be modified to take discrepancies among data sources into account, for example through modelling habitat suitability in N different data sources as the result of a binomial trial with N draws. Whenever (or wherever) sources are judged as differing in quality, their contributions to each pixel's 'occurrence score' could be weighted by their trustworthiness. SDMs can also help target which regions most need sampling effort by looking at the geographic distribution of the discrepancies between models calibrated using different sources of data. Note that these may differ from the regions where sources of data diverge (Fig. 5).

Other possible solutions

The issue of environmental data quality has become more relevant over the last few years with the emergence of environmental Open Data (Reichman *et al.*, 2011). Discrepancies among data sources not only affect species occurrence data: large differences among sources were recently noted for several traits of butterfly species (Fitzsimmons, 2013). These and our results call into question the reliability of research relying on a single source of data; and advocate for an urgent need to develop high-quality databases gathering occurrences (and trait values) of currently existing, naturally regenerating biodiversity. Ecologists have striven for the last 15 years to develop an impressive array of always more sophisticated SDMs. Ironically, they are now in the awkward position where those models cannot provide accurate forecast of changes in species distribution simply because they do not have accurate species distribution data to calibrate and validate them.

The most straightforward way of acquiring such data in a reasonable timeframe may involve participatory science programmes which are drawing increasing attention from scientists, especially in the field of environmental sciences (Dickinson *et al.*, 2012). High-quality databases of species

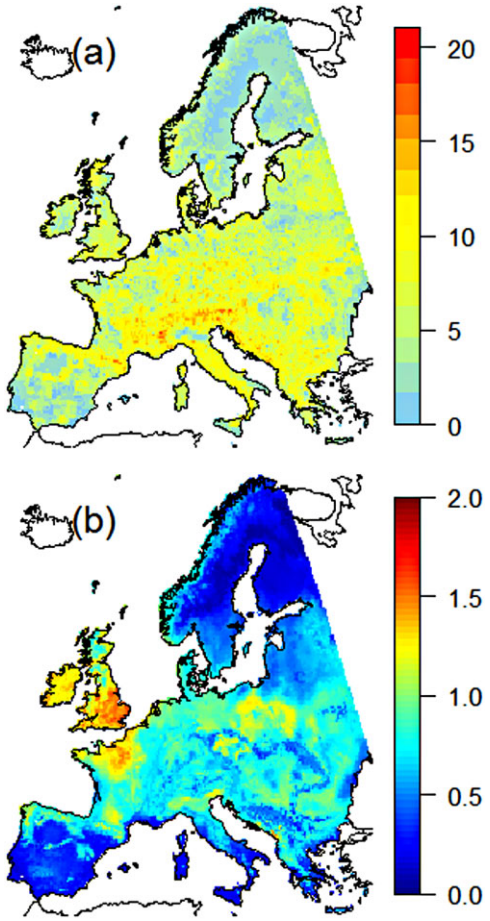

Figure 5 (a) Discrepancies among sources of occurrence data and (b) resulting variance in modelled species distributions (Lambert azimuthal equal area projection). (a) For each pixel the numbers of species (among the 21) for which atlas data diverge. (b) The distribution of variance in predicted current distribution among models generated with BIOMOD, for the same species but from different data sources, summed over all 21 species. Regions of highest among-model variance (b) can differ from those of higher variance in occurrence data (a).

occurrence and reproductive state could be easily obtained through smartphone applications, ensuring easy geolocation, associated with pictures enabling cross-validation of a species' identification or reproductive status. Such programmes have now proven their authority in terms of data collection and scientific input (Devictor *et al.*, 2010; Dickinson *et al.*, 2010; Hochachka *et al.*, 2012). Citizen science programmes per se can be set up for well-known groups of species, for which visual species recognition software already exists (e.g. http://www.leafsnap.com for trees of eastern North America). Nevertheless, citizen science programmes also have the potential to involve self-educated naturalist experts able to identify rarer or difficult taxa.

These efforts could be combined with emerging remote sensing techniques, such as the use of small transmitters able to record migrations with a precision down to a few metres of animals as small as butterflies (e.g. the ICARUS initiative; http://icarusinitiative.org), the use of airborne LiDAR (Korpela *et al.*, 2010) or hyperspectral sensors (Kamaruzaman & Kasawani, 2009) to identify tree species. Data collected by these emerging monitoring techniques should be made publicly available, using common standards of data quality and interoperability (Scholes *et al.*, 2008).

Even with such powerful data collection systems, acquiring accurate species distribution data will take time. In the meantime, SDMs could be modified to take discrepancies among data sources into account (see above), and to deal with 'ensemble datasets' of occurrence data. In the last decade, tremendous efforts have been made to develop ever more sophisticated SDMs. The evaluation of these models, either correlative or process-based, relies heavily on species occurrence data. Accurate occurrence data are an essential prerequisite to achieve the robust forecasting of species distributions and larger-scale biodiversity patterns that stakeholders and policy makers expect, and together with environmental scientists they should urgently realize that such data still need to be collected.

ACKNOWLEDGEMENTS

This research was funded by ANR EVORANGE (ANR-09-PEXT-01102) and SCION (ANR-09-PEXT-01105). A.D. was supported by the ANR EVORANGE. The authors thank F. Massol and D. McKey for advice, E. S. Gritti for providing projections by the model LPJ, and A. Hampe, C. Randin and an anonymous referee for their useful comments that greatly improved the first version of this manuscript.

REFERENCES

Araújo, M.B., Alagador, D., Cabeza, M., Nogués-Bravo, D. & Thuiller, W. (2011) Climate change threatens European conservation areas. *Ecology Letters*, **14**, 484–492.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Beck, J., Ballesteros-Mejia, L., Nagel, P. & Kitching, I.J. (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions*, **19**, 1043–1050.

Bohn, U., Gollub, G., Hettwer, C., Neuhäuslová, Z., Raus, T., Schlüter, H. & Gis, H.W. (2004) Karte der natürlichen Vegetation Europas/Map of the Natural Vegetation of Europe. Interaktive/Interactive CD-ROM – Erläuterungstext, Legende, Karten/Explanatory Text, Legend, Maps. Landwirtschaftsverlag, Münster.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, **15**, 584–593.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.

Cheaib, A., Badeau, V., Boe, J., Chuine, I., Delire, C., Dufrêne, E., François, C., Gritti, E.S., Legay, M., Pagé, C., Thuiller, W., Viovy, N. & Leadley, P. (2012) Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty. *Ecology Letters*, **15**, 533–544.

Chuine, I. & Beaubien, E.G. (2001) Phenology is a major determinant of tree species range. *Ecology Letters*, **4**, 500–510.

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, **16**, 354–362.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. (2010) Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution and Systematics*, **41**, 149–172.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, **10**, 291–297.

Dormann, C.F. (2007) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, **8**, 387–397.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Fitzsimmons, J.M. (2013) How consistent are trait data between sources? A quantitative assessment. *Oikos*, **122**, 1350–1356.

Gritti, E.S., Duputié, A., Massol, F. & Chuine, I. (2013) Estimating consensus and associated uncertainty between inherently different species distribution models. *Methods in Ecology and Evolution*, **4**, 442–452.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. & Peterson, A.T. (2007) What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? *Ecological Monographs*, **77**, 615–630.

Hickler, T., Vohland, K., Feehan, J., Miller, P., Smith, B., Costa, L., Giesecke, T., Fronzek, S., Carter, T.R., Cramer, W., Kühn, I. & Sykes, M.T. (2012) Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global Ecology and Biogeography*, **21**, 50–63.

Higgins, S.I., O'Hara, R.B., Bykova, O., Cramer, M.D., Chuine, I., Gerstner, E.-M., Hickler, T., Morin, X., Kearney, M.R., Midgley, G.F. & Scheiter, S. (2012) A physiological analogy of the niche for projecting the potential distribution of plants. *Journal of Biogeography*, **39**, 2132–2145.

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.-K. & Kelling, S. (2012) Data-intensive science applied to broad-scale citizen science. *Trends in Ecology and Evolution*, **27**, 130–137.

Hortal, J., Lobo, J. & Jiménez-Valverde, A. (2012) Basic questions in biogeography and the (lack of) simplicity of species distributions: putting species distribution models in the right place. *Natureza & Conservação*, **10**, 108–118.

Jalas, J. & Suominen, J. (1964–2010) *Atlas florae Europaeae*. Committee for Mapping the Flora of Europe and Societas Biologica Fennica Vanamo, Helsinki, Finland.

Jiménez-Valverde, A. & Lobo, J.M. (2007) Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, **31**, 361–369.

Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, **14**, 885–890.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kamaruzaman, J. & Kasawani, I. (2009) Hyperspectral remote sensing for tropical rain forest. *American Journal of Applied Sciences*, **6**, 2001–2005.

Korpela, I., Ørka, H.O., Maltamo, M., Tokola, T. & Hyyppä, J. (2010) Tree species classification using airborne LiDAR – effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. *Silva Fennica*, **44**, 319–339.

Kramer, K., Degen, B., Buschbom, J., Hickler, T., Thuiller, W., Sykes, M.T. & de Winter, W. (2010) Modelling exploration of the future of European beech (*Fagus sylvatica* L.) under climate change – range, abundance, genetic diversity and adaptive response. *Forest Ecology and Management*, **259**, 2213–2222.

Lahti, T. & Lampinen, R. (1999) From dot maps to bitmaps: *Atlas Florae Europaeae* goes digital. *Acta Botanica Fennica*, **162**, 5–9.

Lobo, J. (2008) More complex distribution models or more representative data? *Biodiversity Informatics*, **82**, 14–19.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

Meier, E.S., Lischke, H., Schmatz, D.R. & Zimmermann, N.E. (2012) Climate, competition and connectivity affect future migration and ranges of European trees. *Global Ecology and Biogeography*, **21**, 164–178.

Morin, X. & Chuine, I. (2005) Sensitivity analysis of the tree distribution model Phenofit to climatic input characteristics: implications for climate impact assessment. *Global Change Biology*, **11**, 1493–1503.

Morin, X., Viner, D. & Chuine, I. (2008) Tree species range shifts at a continental scale: new predictive insights from a process-based model. *Journal of Ecology*, **96**, 784–794.

Pereira, H.M., Leadley, P.W., Proença, V. *et al.* (2010) Scenarios for global biodiversity in the 21st century. *Science*, **330**, 1496–1501.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martinez-Meyer, E., Nakamura, M. & Araújo, M. (2011) *Ecological niches and geographic distributions*. Princeton University Press, Princeton, NJ.

Pineda, E. & Lobo, J.M. (2012) The performance of range maps and species distribution models representing the geographic variation of species richness at different resolutions. *Global Ecology and Biogeography*, **21**, 935–944.

Randin, C.F., Engler, R., Normand, S., Zappa, M., Zimmermann, N.E., Pearman, P.B., Vittoz, P., Thuiller, W. & Guisan, A. (2009) Climate change and plant distribution: local models predict high-elevation persistence. *Global Change Biology*, **15**, 1557–1569.

Randin, C.F., Paulsen, J., Vitasse, Y., Kollas, C., Wohlgemuth, T., Zimmermann, N.E. & Körner, C. (2013) Do the elevational limits of deciduous tree species match their thermal latitudinal limits? *Global Ecology and Biogeography*, **22**, 913–923.

Real, R., Barbosa, A.M. & Vargas, J.M. (2006) Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, **13**, 237–245.

Reese, G.C., Wilson, K.R., Hoeting, J.A. & Flather, C.H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554–564.

Reichman, O.J., Jones, M.B. & Schildhauer, M.P. (2011) Challenges and opportunities of open data in ecology. *Science*, **331**, 703–705.

Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.

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.

Rondinini, C., Wilson, K.A., Boitani, L., Grantham, H. & Possingham, H.P. (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, **9**, 1136–1145.

Scholes, R.J., Mace, G.M., Turner, W., Geller, G.N., Jurgens, N., Larigauderie, A., Muchoney, D., Walther, B.A. & Mooney, H.A. (2008) Toward a global biodiversity observing system. *Science*, **321**, 1044–1045.

Sitch, S., Smith, B., Prentice, I.C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J.O., Levis, S., Lucht, W., Sykes, M.T., Thonicke, K. & Venevsky, S. (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, **9**, 161–185.

Smith, B., Prentice, I.C. & Sykes, M.T. (2001) Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. *Global Ecology and Biogeography*, **10**, 621–637.

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Swets, J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Thuiller, W., Lavergne, S., Roquet, C., Boulangeat, I., Lafourcade, B. & Araujo, M.B. (2011) Consequences of climate change on the tree of life in Europe. *Nature*, **470**, 531–534.

VanDerWal, J., Murphy, H.T., Kutt, A.S., Perkins, G.C., Bateman, B.L., Perry, J.J. & Reside, A.E. (2013) Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change. *Nature Climate Change*, **3**, 239–243.

Veloz, S.D., Williams, J.W., Blois, J.L., He, F., Otto-Bliesner, B. & Liu, Z. (2012) No-analog climates and shifting realized niches during the late Quaternary: implications for 21st-century predictions by species distribution models. *Global Change Biology*, **18**, 1698–1713.

Wisz, M.S. & Guisan, A. (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, **9**, 8.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. & Culham, A. (2007) How global is the global biodiversity information facility? *PLoS ONE*, **2**, e1124.

Zhu, K., Woodall, C.W. & Clark, J.S. (2012) Failure to migrate: lack of tree range expansion in response to climate change. *Global Change Biology*, **18**, 1042–1052.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Appendix S1 Rates of false positives and false negatives, and overall matches of atlas-derived occurrences, as compared to forest inventory data (ICP dataset).

Appendix S2 Upscaling and downscaling procedure for each of the five sources of distribution data.

Appendix S3 Areas of occurrence of the 21 species according to the available sources of data, and modelled areas for the current period and under scenarios, for models built using occurrence data from each data source.

Appendix S4 Maps showing the number of databases indicating each of the 21 species' occurrences across Europe.

Appendix S5 Maps showing discrepancies between the three atlases.

Appendix S6 Maps showing observed occurrences, modelled current and forecast probabilities of occurrence for the 21 species.

Appendix S7 Proportion of area where models disagree, within the area predicted as suitable by at least one model.

Appendix S8 Post-hoc validation score of two process-based models, using different sources of occurrence as reference.

Appendix S9 'Suitable' area of three European species, as projected by two process-based models as a function of the data source used to define a presence/absence threshold.

BIOSKETCHES

Anne Duputié is an evolutionary ecologist interested in niche evolution, and especially in incorporating microevolutionary processes into process-based SDMs.

Niklaus E. Zimmermann is an ecologist interested in macroecology, ecological theory and niche evolution.

Isabelle Chuine is an ecologist interested in modelling plant niches using traits, and especially phenological traits.

Author contributions: A.D., N.E.Z. and I.C. designed the study. N.E.Z. and I.C. provided data. A.D. performed the analyses and wrote the first version of the manuscript, I.C. and N.E.Z. improved the subsequent versions of the manuscript.

Editor: Arndt Hampe