Title: Improving species distribution models for invasive non-native species with biologically-informed pseudo-absence selection

Running title: Invasive species distribution models

Authors: Daniel Chapman[1], Oliver L. Pescott[2], Helen E. Roy[2], Rob Tanner[3]

Institutional affilitations:

1 UKRI Centre for Ecology & Hydrology, Edinburgh EH26 0QB, UK

2 UKRI Centre for Ecology & Hydrology, Wallingford OX10 8BB, UK

3 European and Mediterranean Plant Protection Organisation, 21 Boulevard Richard Lenoir, 75011 Paris, France

Corresponding author: Daniel Chapman

Email addresses: Daniel Chapman dcha@ceh.ac.uk, Oliver L. Pescott olipes@ceh.ac.uk, Helen Roy hele@ceh.ac.uk, Rob Tanner rt@eppo.int

Acknowledgements: This research was funded by European Union Life Programme Preparatory project LIFE15 PRE/FR/000001. We thank the Expert Working Groups who performed EPPO Pest Risk Analyses for the five study species and provided us with data and species information to build our models.

Abstract

Aim: We present a novel strategy for species distribution models (SDMs) aimed at predicting the potential distributions of range-expanding invasive non-native species (INNS). The strategy combines two established perspectives on defining the background region for sampling 'pseudo-absences' that have hitherto only been applied separately. These are the accessible area, which accounts for dispersal constraints, and the area outside the environmental range of the species and therefore assumed to be unsuitable for the species. We tested an approach to combine these by fitting SDMs using background samples (pseudo-absences) from both types of background.

Location: Global

Taxon: Invasive non-native plants: *Humulus scandens*, *Lygodium japonicum*, *Lespedeza cuneata*, *Triadica sebifera*, *Cinnamomum camphora*

Methods: Presence-background (or presence-only) SDMs were developed for the potential global distributions of five plant species native to Asia, invasive elsewhere and prioritised for risk assessment as emerging INNS in Europe. Models where 'pseudo-absences' were selected only from the accessible background were compared to models based on accessible and unsuitable domains, with the latter defined using biological knowledge of the species' key limiting factors.

Results: Combining the unsuitable and accessible backgrounds expanded the range of environments available for model fitting and caused biological knowledge about ecological unsuitability to influence the fitted species-environment relationships. This improved the realism and accuracy of distribution projections globally and within the species' ranges.

Main conclusions: Correlative SDMs remain valuable for INNS risk mapping and management, but are often criticised for a lack of biological underpinning. Our approach partly addresses this criticism by using prior knowledge of species' requirements or tolerances to define the unsuitable background for modelling, while also accommodating dispersal constraints through considerations of accessibility. It can be implemented with current SDM software and results in more accurate and realistic distribution projections. As such, wider adoption has potential to improve SDMs that support INNS risk assessment.

Keywords: Biomod; climate envelope; ecological niche model; invasive alien species; Maxent; pest risk assessment; presence-absence; presence-only; presence-background; pseudo-absence.

Introduction

Human transport of species beyond their native ranges, leading to biological invasions, is an important driver of ecological change, impacting biodiversity and ecosystem function (Vilà et al., 2011). Decision making about the control and management of invasive non-native species (INNS) is often underpinned by scientific risk assessments, and species distribution models (SDM) are increasingly seen as a valuable tool for this (Jeschke & Strayer, 2008; Václavík & Meentemeyer, 2009; Jiménez-Valverde et al., 2011). The purpose of SDMs applied in this context is to generate risk maps that predict the potential distribution of an INNS as a function of climate and other environmental gradients (Jiménez-Valverde et al., 2011). Specifically, these represent the relative likelihood of establishment should the species be introduced or disperse to each location in the modelled landscape (Elith, 2013). Risk maps can be used for prioritisation of surveillance and management (Peterson & Robins, 2003; Gormley et al., 2011), to estimate the potential spread of emerging INNS in current and future climates (Jiménez-Valverde et al., 2011; Branquart et al., 2016) and to understand the biological and anthropogenic mechanisms governing invasions (Broennimann et al., 2007; Chapman et al., 2014, 2017; Storkey et al., 2014). Clearly, there is a need for robust and accessible SDM tools and methods to ensure the most accurate possible estimation of the potential distributions of INNS.

Species prioritised for risk assessment in one area have typically already established invasive non-native distributions in other parts of the world (Roy et al., 2014; Branquart et al., 2016; Tanner et al., 2017) snecessitating global-scale models and the pooling of distribution data from native and already-invaded ranges (Broennimann & Guisan, 2008; Mainali et al., 2015). Unfortunately species' distributions are rarely documented comprehensively at the spatial resolutions of SDMs (Boakes et al., 2010). Therefore, global-scale models are typically developed using statistical algorithms that contrast the environmental conditions where the species is known to occur with those at 'pseudo-absence' locations sampled from a background domain specified by the modeller. Such SDMs are often referred to as presence-only models (Pearce & Boyce, 2006) but we use the term presence-background to differentiate them from 'one-case' or true presence-only models that use only the species presences and not the background (Guillera-Arroita et al., 2015). We also differentiate the ‘pseudo-absence’-based presence-background models that are the focus of this study from point process models for species distributions (Warton & Shepherd, 2010). Point process models generalise presence-background models on a more formal statistical basis. However, to our knowledge they are not suitable for grid cell-resolution distribution data, have not been applied for global-scale modelling of INNS and are far less commonly used than well-known presence-background models such as Maxent (Phillips et al., 2008) or the regression and machine learning approaches implemented through software platforms such as Biomod (Thuiller et al., 2009, 2016).

One important issue when fitting presence-background models to INNS distribution data is that their global distributions are by definition in a non-equilibrium state and are structured by both the species' environmental tolerances and natural and anthropogenic dispersal constraints (Václavík & Meentemeyer, 2009; Gallien et al., 2010; Chapman et al., 2016). As a consequence, there are suitable but unoccupied regions in which climatic and environmental conditions would permit establishment by the species, but where invasion has not been realised through dispersal. If such regions are included in the background domain, then the model will conflate lack of presence of the species due to dispersal constraints with a lack of presence due to environmental unsuitability, potentially biasing the species-environment relationships and the prediction of potential distributions. Current approaches to reduce this bias emphasise restricting the background domain to an ‘accessible area’ within dispersal range of the occurrences (Barve et al., 2011; Elith, 2013; Mainali et al., 2015). Although likely to lessen dispersal biases in presence-background models, we suggest this may be overly restrictive for modelling aimed at risk mapping. If background samples are only drawn in close proximity to the occurrences then the range of environmental conditions used to train the model may be insufficient to fully characterise species-environment relationships, impeding the transfer of predictions into other regions (Thuiller et al., 2004; Fitzpatrick & Hargrove, 2009).

Here, we propose a biologically-informed approach to improve presence-background models for highly dispersal-limited species, such as those undergoing invasive range expansion. The goal is to exclude suitable but unoccupied regions while also maximising the range of environmental conditions used to train the model. As such, we combine two familiar types of background domain – an accessible background in proximity to species' occurrences (Barve et al., 2011; Mainali et al., 2015) and an unsuitable background outside the environmental envelope of the species (Thuiller et al., 2004; Chefaoui & Lobo, 2007; Le Maitre et al., 2008). Those previous studies have tested both types of background in isolation, but the novel contributions of this study are to combine both types of background, and to emphasise the definition of the unsuitable background using biological knowledge of key limiting factors for the species, e.g. places that do not reach minimum growing temperatures or exceed maximum drought tolerance. By modelling the global distributions of five invasive non-native plants we demonstrate that this constrains the presence-background models to fit more biologically plausible response functions and increases the accuracy of distribution projections.

Methods

Overview

Our aim was to compare global-scale presence-background SDMs for INNS developed using background domains defined in the standard way (as only the accessible region sensu Barve et al., 2011) or through our proposed new approach of combining accessible and unsuitable background regions (Figure 1-2). Models were developed to predict the potential distributions of five plant species that are native to temperate and tropical east Asia, highly invasive in other parts of the world and have been prioritised for risk assessment as potentially-emerging invasive non-native plant species in Europe (Branquart et al., 2016; Tanner et al., 2017). The species represent a range of life histories including an annual climbing vine (*Humulus scandens*), a perennial climbing fern (*Lygodium japonicum*), a perennial semi-woody forb (*Lespedeza cuneata),* a deciduous tree (*Triadica sebifera*) and an evergreen tree (*Cinnamomum camphora*).

Data for modelling

Species occurrences were obtained from a range of sources including Global Biodiversity Information Facility (GBIF), USGS Biodiversity Information Serving Our Nation (BISON), Integrated Digitized Biocollections (iDigBio), iNaturalist, Early Detection and Distribution Mapping System (EDDMapS) and from the members of the European and Mediterranean Plant Protection Organisation (EPPO) expert working groups conducting Pest Risk Analyses for the region. With these experts, we scrutinised occurrence records and removed any that appeared dubious, casual or cultivated (e.g. botanic gardens) or where the georeferencing was too imprecise (e.g. country or island centroids). The remaining records were gridded at a 0.25 x 0.25 degree resolution for global modelling. As a proxy for plant recording effort, the total number of vascular plant records (phylum Tracheophyta) per grid cell was also obtained from GBIF (see Appendix S1 in Supporting Information).

Three predictor variables, derived from WorldClim v1.4 (Hijmans et al., 2005), were selected to represent basic constraints on plant distributions. These were mean temperature of the warmest quarter (Bio10, °C) reflecting the growing season thermal regime, mean minimum temperature of the coldest month (Bio6, °C) reflecting exposure to winter cold and the climatic moisture index (CMI, ratio of annual precipitation, Bio12, to potential evapotranspiration, then ln + 1 transformed) reflecting drought stress. Potential evapotranspiration was estimated following Zomer et al. (2008).

Definition of the background domains

Background samples (pseudo-absences) were drawn from two distinct regions – an accessible region and a region considered to be environmentally unsuitable for the species based on knowledge of its tolerances or requirements (Figures 1 and 2). Though both types of background represent established concepts within distribution modelling, to our knowledge, this is the first study to test whether modelling is improved by combining both types of background domain.

The accessible background attempts to cover only the region where the species has had opportunity to disperse and sample the environment (Thuiller et al., 2004; VanDerWal et al., 2009; Barve et al., 2011;

Mainali et al., 2015). It has generally been defined as a zone around the occurrence data, which could be selected statistically or informed by dispersal abilities of the species (Elith, 2013; Senay et al., 2013). For invasive non-native species, the size of the accessible region will generally be more limited in the invaded range than the native one, assuming stronger dispersal constraints associated with shorter residence time (Mainali et al., 2015). In our application, we defined the native accessible areas using a 400 km geodesic buffer around the minimum convex polygon bounding all native occurrences (Figure 1a). In the non-native region, we used a conservative 4-cell neighbourhood around each occurrence grid cell, equivalent to a ~30 km buffer (Figure 1b). Though somewhat arbitrary, these buffer sizes are consistent with ones performing well in other presence-background SDM studies (VanDerWal et al., 2009; Mainali et al., 2015), and a sensitivity analysis showed model outputs were not strongly influenced by the choice of native buffer size (see Appendix S5).

The unsuitable background concept originates from existing ideas about sampling pseudo-absences only outside of the environmental envelope in which species' presences are found (Thuiller et al., 2004; Chefaoui & Lobo, 2007; Le Maitre et al., 2008; Senay et al., 2013). The rationale is to produce training datasets that maximise the distinctiveness of suitable environmental conditions from the background and therefore boost the model discrimination. However, it may also reduce model accuracy within the environmental and geographical range of the species (Acevedo et al., 2012). These previous studies simply screened out the ranges of all environmental variables at presence locations, or used preliminary modelling to determine unsuitable regions. However, in this study we instead used prior biological knowledge and expert opinion about the species' limiting factors to define the unsuitable conditions (Figures 1 and 2) in the expectation that this biological information would be captured in the fitted species-environment relationships. Appropriate rules to define unsuitability were determined in consultation with species experts participating in their EPPO expert working groups. Their expert judgement informed us on the type of limit deemed to be most important for the species in different parts of its range (e.g. summer cold, drought), followed by identification of key thresholds from the literature and comparison with extreme values at the occurrence locations of the species (see Appendix S2).

Sampling from the background domain

To combine both types of backgrounds, we obtained background samples from both the accessible region and from the unsuitable region outside of the accessible region (Figures 1-2). The effect was therefore to exclude potentially suitable but inaccessible regions from the combined background sample. To reduce sampling variation, ten replicate background samples were generated. Presence-background models were developed for each background sample and then their predictions were averaged.

The accessible region was sampled using target group sampling to reduce bias in the observed distribution due to spatial sampling effort variation (Phillips, 2009; Ranc et al., 2017). This involves weighting the background sampling by the recording density of a broader taxonomic group, which is assumed to represent recording bias for the focal species. In our modelling we used the GBIF record density of vascular plants (Tracheophyta) as a target group to weight background sampling. From the accessible region we drew the same number of background samples as there were occurrences, weighted by the vascular plant record density as a target group. This ensured that the accessible area background sample contained the same degree of recording bias as the occurrence data, assuming the proxy for recording effort was appropriate.

The unsuitable region was sampled with simple random sampling because we considered that recording bias should not be a relevant consideration in the observed lack of presence from environments in which the species cannot occur. In other words, we were confident of absence in the unsuitable regions. Although we could have nevertheless applied target group sampling, random sampling has the potential advantage of accumulating background samples from environments where there is little survey effort (e.g. very cold conditions), resulting in the widest range of environments from which to model species-environment relationships. For the five species in this study, 3000 random samples were taken from the unsuitable region, outside the accessible region. A sensitivity analysis on the number of unsuitable background samples showed that the number of sampling points was not critical to model performance (see Appendix S5).

Ensemble presence-background modelling

For each species, presence-background models were developed using background samples from only the accessible area and using the combined background samples from both the accessible and unsuitable area. Ensemble models were fitted using BIOMOD (biomod2 R package v3.3-7) (Thuiller et al., 2009, 2016) using seven statistical algorithms: generalised linear models (GLM) with linear and quadratic terms for each predictor, generalised additive models (GAM) with a maximum of four degrees of freedom per variable, multivariate adaptive regression splines (MARS), generalised boosting models (GBM), random forests (RF), artificial neural networks (ANN) and Maxent (Phillips et al., 2008). These were combined into an ensemble model by scaling their predictions with a binomial GLM and then averaging them weighted by predictive AUC scores (80:20% split for training and evaluation). AUC is commonly used for ensemble model weighting and is the BIOMOD default option (Thuiller et al., 2009, 2016). Although AUC does not provide an objective measure of model performance for presence-only models (Lobo, 2008) it is informative about the relative discrimination abilities of different algorithms evaluated on the same data. It also provides a conservative model weighting scheme, since a perfect model (AUC=1) will have only twice the weight of a random model (AUC=0.5). Therefore, we ensured poorly performing algorithms did not disproportionately affect the weighted average by rejecting them from the ensemble. Rejection was based on modified *z*-scores for their predictive AUC (Crosby, 1993) with algorithms with $z < -1$ being rejected.

The importance of each variable to model fitting was estimated through the BIOMOD default procedure (Thuiller et al., 2009). Species-environment relationships were examined by constructing univariate response curves where predictions of the ensemble model were made while fixing the other variables at typical suitable values (median in the presence grid cells). Global projections of the ensemble models were restricted to where the environmental predictors lay inside the ranges used in model training, avoiding model extrapolation (Fitzpatrick & Hargrove, 2009). Models based only on the accessible background were compared with those based on the combined accessible and unsuitable background in a standardised way. To do this we used AUC to evaluate their discrimination of presences from background samples in both the accessible background domain and in the accessible and unsuitable background domain. As mentioned above, AUC in this context is informative about the relative discrimination power of different model specifications on the same data. By comparing model AUCs within the same background regions we ensured a fair comparison.

Results

Adequate numbers of grid cells with presences were obtained for modelling the five study species (695 for *Cinnamomum camphora*, 754 for *Humulus scandens*, 1723 for *Lespedeza cuneata*, 975 for *Lygodium japonicum* and 855 for *Triadica sebifera*) (see Appendix S2). For every species, models combining the accessible and unsuitable backgrounds discriminated presences more successfully than models using only the accessible background (Table 1 and Appendix S3). Clear improvements in model performance at predicting the global range of the species were obtained (mean AUC improvement of 0.048 across the full model backgrounds). AUC gains for the combined background were small but still appreciable within the accessible region, representing projections within the species' observed ranges. From the binomial distribution, the probability of getting AUC improvements for all five species by chance is $P = 0.063$. Furthermore, in the sensitivity analysis of accessible region size and number of unsuitable background samples (see Appendix S5) the combined background model had higher accessible-region AUC in 36 out of 45 model permutations (80%). This is a significant departure from a 50:50 chance of AUC improvement according to a binomial generalised linear mixed model with random species effect, which yielded a fixed intercept term greater than zero ($P = 0.027$).

Models based on the combined accessible and unsuitable background yielded partial response curves constrained with near zero suitability when conditions exceeded the thresholds used to define the unsuitable region (Figure 3). Models developed using only the accessible background generally gave qualitatively similar response curves, but spanned a narrower range of suitability values and therefore provided a less clear distinction between high and low suitability. Furthermore, there were examples where the response curves from both models differed markedly, most clearly seen in the responses of *Cinnamomum camphora* to moisture (CMI) and of *Lespedeza cuneata* to winter temperature (Bio6) (Figure 3). It can also be seen from Figure 3 that inclusion of the unsuitable background increased the range of predictor gradients available to train the models.

Projections of potential non-native ranges made with both types of model were also qualitatively similar in general (Figures 4 and 5, see Appendix S4 for global and native range projections). However, when the unsuitable background was included then the projections generally made a sharper delineation between very low and high suitability, and projections were not impeded by extrapolation. There were also some notable differences in the details of the projections. For example, in North America the inclusion of the unsuitable region reversed the predictions of suitability for *Cinnamomum camphora*, *Triadica sebifera* and *Lygodium japonicum* invasion in arid parts of south western USA and the prediction of suitability for *Lespedeza cuneata* in north eastern USA and southern Canada (Figure 4). In Europe, where the species are not so well established, the inclusion of the unsuitable background domain produced substantially larger regions predicted to have high suitability for *C. camphora* and *T. sebifera*, and reversed the prediction of suitability for *L. japonicum* in Spain (Figure 5).

Both types of model suggested that the species have reached their niche limits in the native range (see Appendix S4) but are capable of further niche filling and non-native range expansion in North America (Figure 4) and Europe (Figure 5). In Europe, both models predict that *Humulus scandens* and *Lespedeza cuneata* may be able to invade widely in central and northern regions (Figure 5). By contrast, *Cinnamomum camphora*, *Lygodium japonicum* and *Triadica sebifera* may be restricted to southern and relatively frost-free parts of Western Europe.

Discussion

Strategies for selecting background samples or pseudo-absences for presence-background species distribution models have received a great deal of attention (e.g. Thuiller et al., 2004; Chefaoui & Lobo, 2007; VanDerWal et al., 2009; Barve et al., 2011). The novel contribution of this study is to combine two different perspectives on defining the background region that have hitherto been considered separately. These perspectives are the accessible area (Barve et al., 2011) and the area outside the environmental range of the species, and therefore assumed to be unsuitable for the species (Thuiller et al., 2004). Previous work on modelling invasive non-native species has generally either emphasised the usefulness of the former for accommodating dispersal constraints (Mainali et al., 2015) or evaluated the latter as a way of boosting the discrimination between suitable and unsuitable habitat (Le Maitre et al., 2008). To our knowledge, the only previous attempt to jointly consider both perspectives did so in a more limited way than this study, by excluding parts of the accessible region that were outside the environmental range of the species (Senay et al., 2013). Here, we tested a new approach in which separate background samples were obtained from the accessible region, regardless of environmental values, and from an unsuitable region defined using prior biological knowledge. By modelling the global distributions of five invasive non-native plant species we conclude that the new strategy performed better for projection of regional and global potential distributions than when models were fitted with just the accessible region.

This was evidenced by a consistent improvement in model discrimination of presences when the modelling sampled from a biologically-informed unsuitable background. This was most clearly seen across the combined accessible and unsuitable background, suggesting better performance for global projection. However, more interesting was the marginal improvement within the accessible region itself, indicating better discrimination within the observed species' ranges. Our expectation was that discrimination within the range would not be improved by increasing the size of the modelling domain. Indeed, previous studies have found that large geographical background domains increase the power of SDMs to model species' broad geographic ranges but decrease their representation of suitability gradients within the range (Thuiller et al., 2004; VanDerWal et al., 2009). Unlike previous studies, our approach may have resulted in improved performance for both purposes because we explicitly tried to exclude ‘suitable-but-not-reached’ locations from the larger (unsuitable) backgrounds. As such, we suggest that biologically-informed specification of a large modelling domain may reduce the trade-off between prediction of suitability gradients at large and small spatial scales.

The influence of the unsuitable background on species-environment relationships was clearly seen in the response curves and projections of the models. In most cases, response curves were qualitatively similar to those fitted by models based only on the accessible background. However, the inclusion of the unsuitable background had three clear effects. First, it ‘anchored’ the curves by constraining the models to fit near-zero suitability where the climate variables exceeded the thresholds of the species, providing a more pronounced delineation of suitability gradients. Second, the response curves were less complex or multi-modal than those from models using only the accessible background, which is more consistent with niche theory (Austin, 2002). Third, the response curves reflected prior assumptions about environmental limitation of the species and as such were more consistent with ecological understanding of the species. For instance, models for *Cinnamomum camphora*, *Lygodium japonicum* and *Triadica sebifera* estimated a strong limitation by low moisture availability (CMI), precluding potential establishment in arid regions such as south west USA and Spain. This is consistent with empirical demonstrations of water stress reducing growth and survival of these species. For example, shoot growth of *C. camphora* is 30% lower at 40% field water capacity than at 80% (Zhao et al., 2006), water restriction suppresses *T. sebifera* seedling growth by 30-80% (Barrilleaux & Grace, James, 2000) and its seedlings wilt and die in arid western USA unless planted in moist micro-habitats such as river banks (Bower et al., 2009). Similarly, inclusion of the unsuitable region strongly limited suitability of *Lespedeza cuneata* by very cold winters, consistent with known frost sensitivity of the species especially in relation to late spring frosts (Gucker, 2010). The broader conclusion is that sampling from an unsuitable background forces the statistical models to learn species-environment relationships that reflect the prior knowledge of the species’ tolerances or niche requirements used to define the unsuitable domain. As such, we suggest that our approach offers a way of incorporating prior biological knowledge into correlative species distribution models, and as such can address the common criticism that they lack strong biological underpinning (Austin, 2002; Dormann et al., 2011; Chapman et al., 2014).

Sensitivity analyses suggested that our findings were not overly sensitive to the size of the accessible region, number of background samples or precise rules for determining unsuitable conditions (see Appendix S5). However, success of the modelling approach likely relies on careful selection of the appropriate environmental limits to define the unsuitable region in the modelling (Le Maitre et al., 2008). A strength of this study is that it was done in consultation with experts performing risk assessments for invasion of Europe by the species. These experts were able to provide guidance on the key limiting factors relevant for different parts of the invaded and native ranges of the species. Some of the species have been well studied in their other invaded ranges and we were able to draw upon previous experimental studies that had determined tolerance thresholds for the species (see Appendix S2). Where this information was lacking, we used upper or lower bounds on the environmental values at the species presences to define thresholds for modelling. Even where empirical estimates of threshold values were available, we still recommend checking for consistency with environmental values at the distribution data, since species-environment relationships are highly scale-dependent (Siefert et al., 2012) and species can occupy broadly unsuitable regions if suitable micro-habitats are available. Given the reliance on prior studies or expert judgement about species' limiting factors or tolerances, our methods are probably most suitable for relatively well known species and less applicable to species where knowledge of its environmental limits are lacking. However, regional risk assessments for emerging invasive non-native species generally prioritise species that behave invasively in other parts of the world (Roy et al., 2014; Branquart et al., 2016; Tanner et al., 2017) suggesting that our modelling approach might be widely applicable for species of concern.

Risk assessment is a critical tool in the management of emerging invasive non-native species and requires robust prediction of where is vulnerable to ongoing species establishment and spread (Keller et al., 2007;

Jiménez-Valverde et al., 2011). This study shows that defining the model background to accommodate considerations of accessibility as well as prior biological knowledge of environmental unsuitability has the potential to improve global-scale presence-background models for emerging invasive non-native species. The methods developed and tested here are fully implemented by manipulating the model input data, and as such they can be implemented simply using standard presence-background modelling software such as Biomod (Thuiller et al., 2009) or Maxent (Phillips et al., 2008). Furthermore, they result in presence-background models that are more strongly underpinned by biological knowledge rather than being solely driven by distribution data, which are often incomplete and biased. As such, wider adoption of these approaches should improve global-scale modelling of invasive non-native species distributions, contributing to more accurate risk assessment and better management of their impacts.

Tables

Table 1. Discrimination performance of ensemble model projections for the potential global distribution of five plant species developed using two different background region specifications (A = accessible background, AU = accessible and unsuitable background). Discrimination performance is given as AUC (Area Under the receiver-operator Curve) for the combined 10 background samples from only the accessible background region, or for the accessible and unsuitable background region. For presence-only data AUC is the probability that a species presence has a higher projected suitability than a background sample.

Species	**AUC in the accessible background**		**AUC in the accessible and unsuitable background**	
	Accessible model	**Accessible and unsuitable model**	**Accessible model**	**Accessible and unsuitable model**
Cinnamomum camphora	0.691	0.708	0.864	0.982
Humulus scandens	0.786	0.793	0.970	0.984
Lespedeza cuneata	0.9110	0.9113	0.969	0.983
Lygodium japonicum	0.850	0.870	0.929	0.983
Triadica sebifera	0.785	0.789	0.940	0.983

Figures

Figure 1. Illustration of part of the regions from which background samples (pseudo-absences) were drawn for modelling *Humulus scandens*. Dark shading shows the accessible background, where the species is assumed to have had chance to disperse to and sample. Light shading shows the unsuitable background, defined using biological information on the key limiting factors of the species (see Appendix S2). (a) The Asian native range of the species, where accessibility was defined with a buffer around the minimum convex polygon of the occurrences. (b) The North American part of the invaded range, where accessibility was more restricted to represent stronger dispersal constraints during the invasive range expansion.

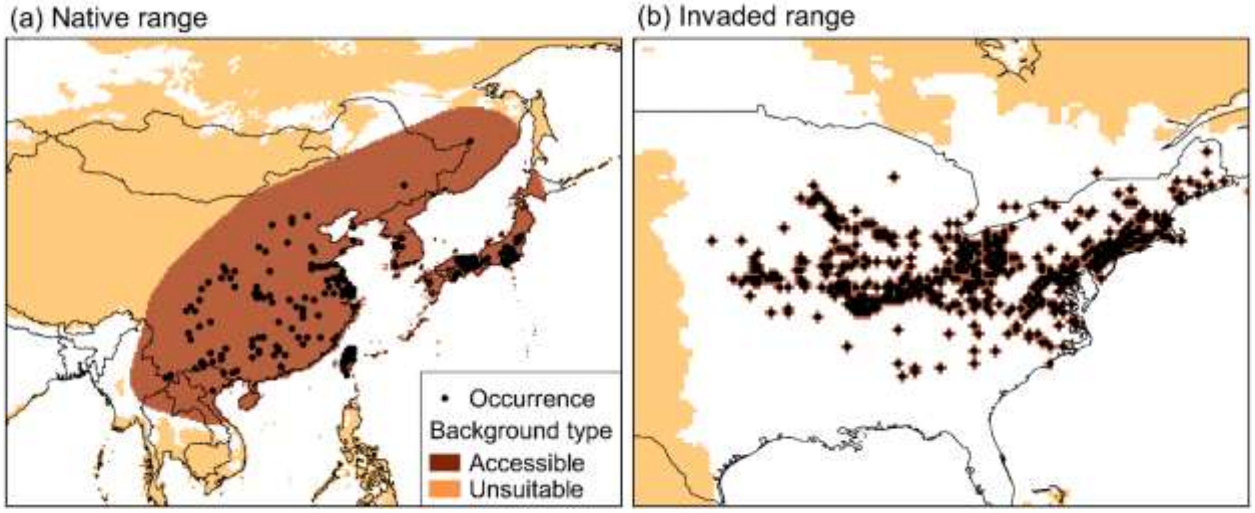


Figure 2. Flow chart for implementing the biologically-informed pseudo-absence selection for presence-background modelling of invasive non-native species.

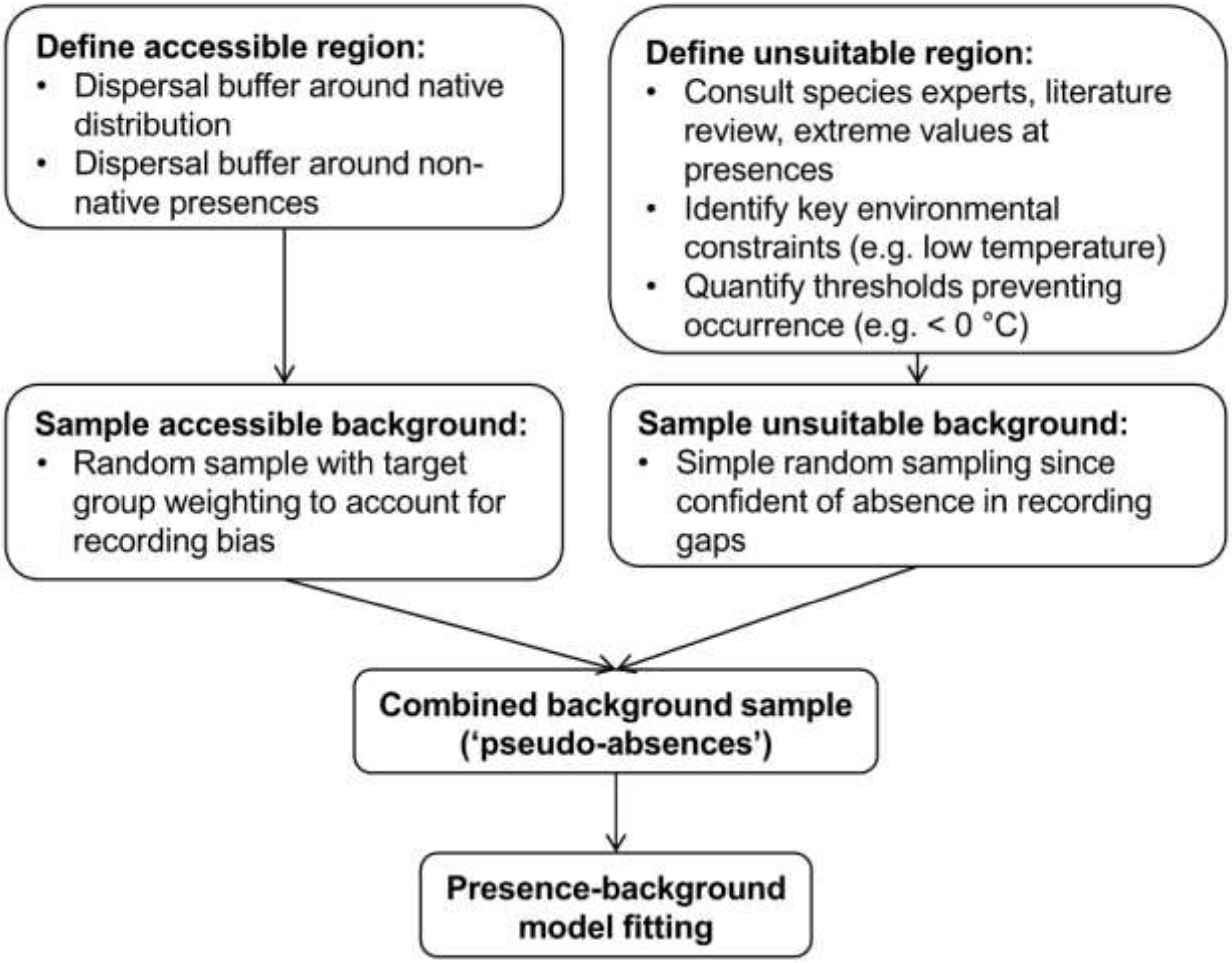

Define accessible region:
• Dispersal buffer around native distribution
• Dispersal buffer around non-native presences
Define unsuitable region:
• Consult species experts, literature review, extreme values at presences
• Identify key environmental constraints (e.g. low temperature)
• Quantify thresholds preventing occurrence (e.g. < 0 °C)
Sample accessible background:
• Random sample with target group weighting to account for recording bias
Sample unsuitable background:
• Simple random sampling since confident of absence in recording gaps
Combined background sample ('pseudo-absences')
Presence-background model fitting

Figure 3. Partial response plots fitted by the ensemble models showing the predicted suitability when other variables are fixed at suitable values for the species (medians in the presence grid cells). Curves span the range of the variables in the training data. Curve colour differentiates the models with background domains based only on the accessible region and those including the unsuitable region. Variable codes: Bio6 = mean minimum temperature of the coldest month (°C); Bio10 = mean temperature of the warmest quarter (°C); CMI = climatic moisture index (ln+1 transformed).

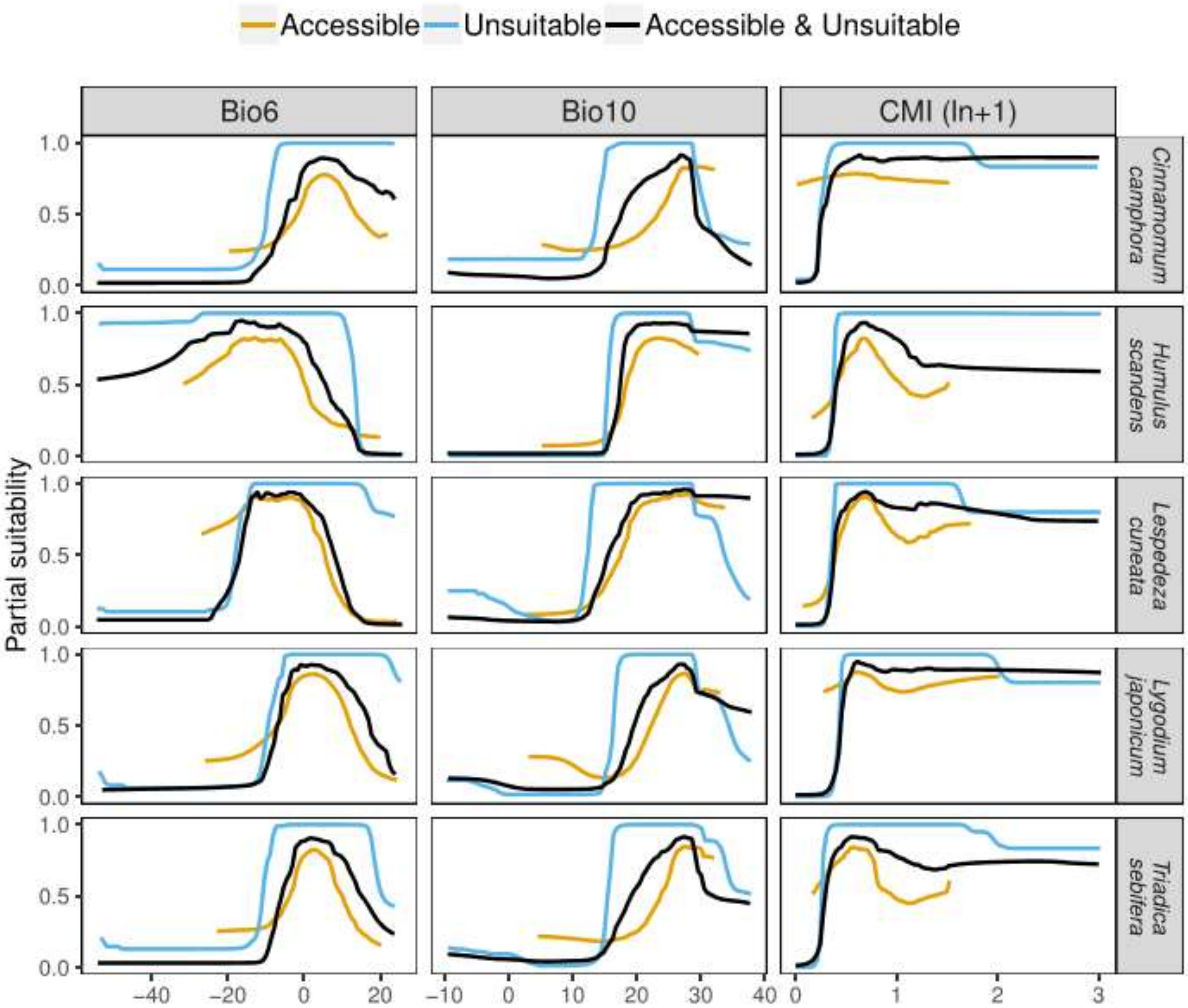


Figure 4. Potential non-native distributions of five Asian plant species in the USA, where all are already established invasive non-native species with expanding ranges. Projections are from models where the background domain is either just the accessible area, or the accessible and unsuitable region. Points show the occurrences and shading indicates the predicted suitability. Blank land areas are where the model could not project suitability because one or more predictors was outside the range of the training data.

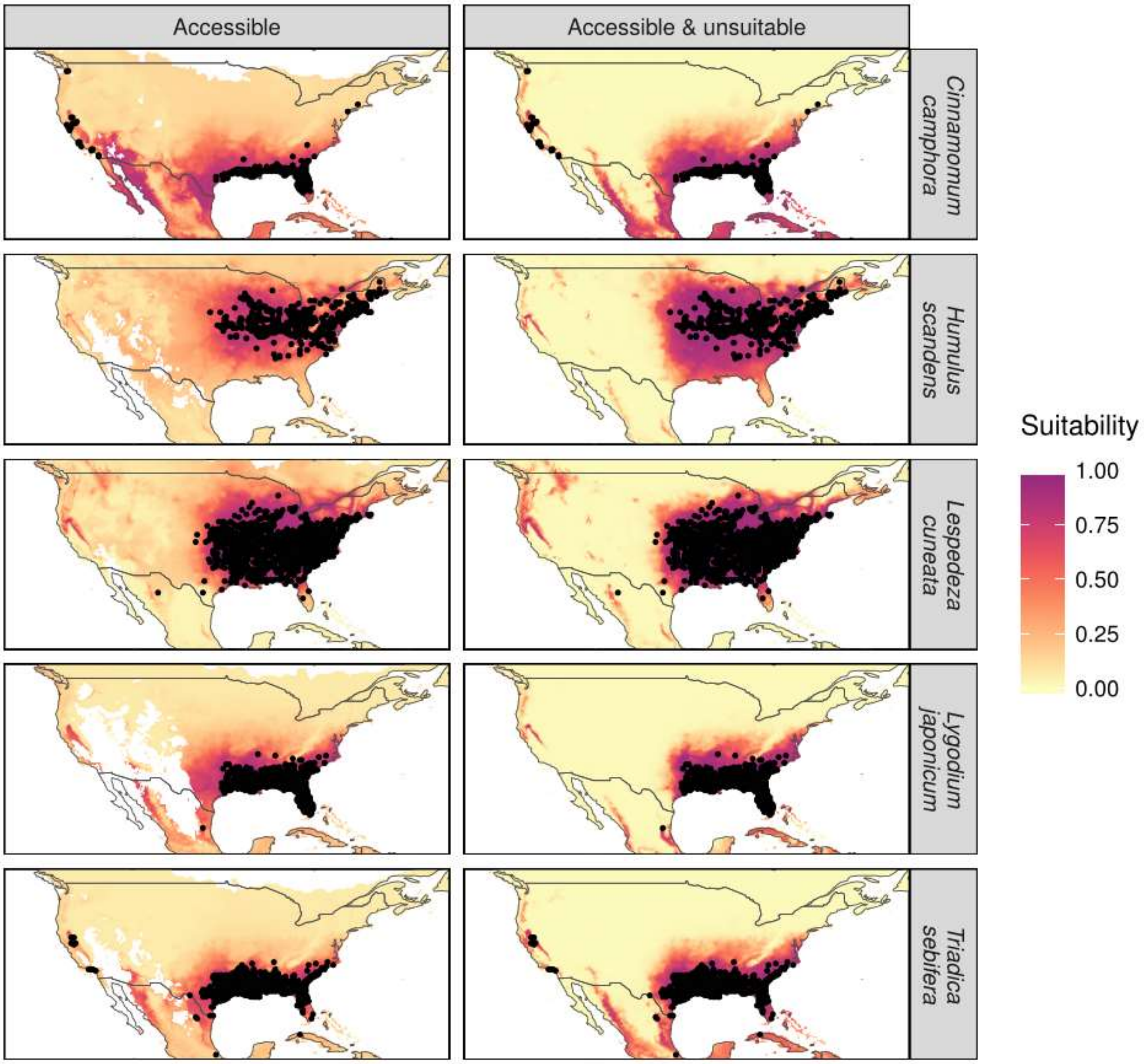
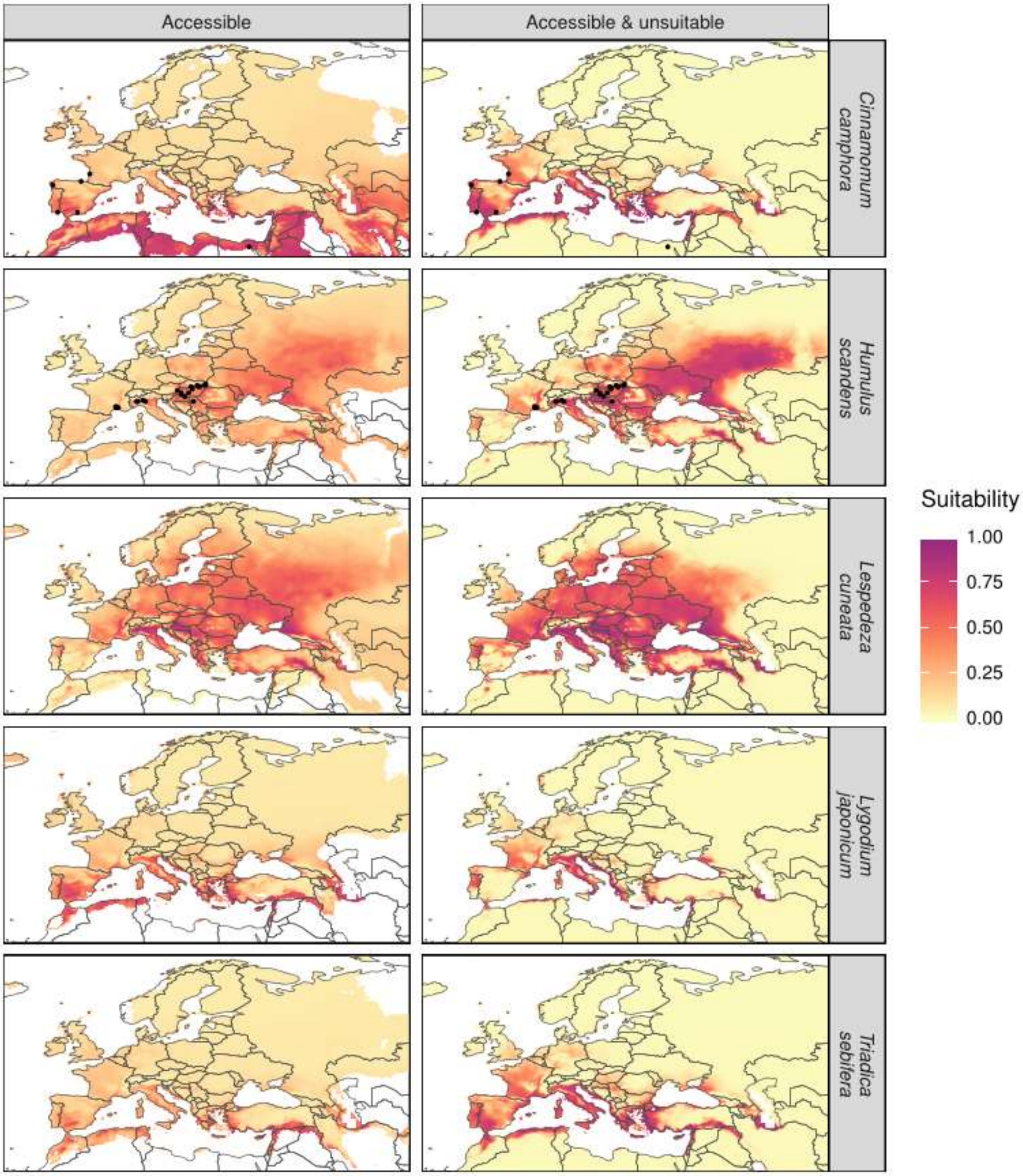


Figure 5. Potential distributions of five Asian plant species in Europe, where the species are currently absent or emerging invasive non-native species, equivalent to Figure 4.



References

Acevedo, P., Jiménez-Valverde, A., Lobo, J.M., & Real, R. (2012) Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, **39**, 1383–1390.

Austin, M. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modeling*, **157**, 101–118.

Barrilleaux, T.C. & Grace, James, B. (2000) Growth and invasive potential of Sapium sebiferum (Euphorbiaceae) within the coastal prairie region: the effects of soil and moisture regime. *American Journal of Botany*, **87**, 1099–1106.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., & Villalobos, F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, **222**, 1810–1819.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-Qing, D., Clark, N.E., O’Connor, K., & Mace, G.M. (2010) Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385.

Bower, M.J., Aslan, C.E., & Rejmánek, M. (2009) Invasion potential of Chinese tallowtree (*Triadica sebifera*) in California’s Central Valley. *Invasive Plant Science and Management*, **2**, 386–395.

Branquart, E., Brundu, G., Buholzer, S., Chapman, D., Ehret, P., Fried, G., Starfinger, U., van Valkenburg, J., & Tanner, R. (2016) A prioritization process for invasive alien plant species incorporating the requirements of EU Regulation no. 1143/2014. *EPPO Bulletin*, **46**, 603–617.

Broennimann, O. & Guisan, A. (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, **4**, 585–589.

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A.T., & Guisan, A. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 701–709.

Chapman, D.S., Haynes, T., Beal, S., Essl, F., & Bullock, J.M. (2014) Phenology predicts the native and invasive range limits of common ragweed. *Global Change Biology*, **20**, 192–202.

Chapman, D.S., Makra, L., Albertini, R., Bonini, M., Páldy, A., Rodinkova, V., Šikoparija, B., Weryszko-Chmielewska, E., & Bullock, J.M. (2016) Modelling the introduction and spread of non-native species: international trade and climate change drive ragweed invasion. *Global change biology*, **22**, 3067–3079.

Chapman, D.S., Scalone, R., Štefanić, E., & Bullock, J.M. (2017) Mechanistic species distribution modeling reveals a niche shift during invasion. *Ecology*, **98**, 1671–1680.

Chefaoui, R.M. & Lobo, J.M. (2007) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478–486.

Crosby, T. (1993) *How to Detect and Handle Outliers.* ASOC Quality Press, Milwaukee.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., & Singer, A. (2011) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Elith, J. (2013) Predicting distributions of invasive species. *Invasive Species: Risk Assessment and Management* (ed. by A.P. Robinson, T. Walshe, M.A. Burgman, and M. Nunn), pp. 93–129. Cambridge University Press, Cambridge, UK.

Fitzpatrick, M.C. & Hargrove, W.W. (2009) The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, **18**, 2255–2261.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I., & Thuiller, W. (2010) Predicting potential distributions of invasive species: where to go from here? *Diversity and Distributions*, **16**, 331–342.

Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S.L., Scroggie, M.P., & Woodford, L. (2011) Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology*, **48**, 25–34.

Gucker, C. (2010) Available at: http://www.fs.fed.us/database/feis/.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., Mccarthy, M.A., Tingley, R., & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276–292.

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

Jeschke, J.M. & Strayer, D.L. (2008) Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences*, **1134**, 1–24.

Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P., & Lobo, J.M. (2011) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785–2797.

Keller, R.P., Lodge, D.M., & Finnoff, D.C. (2007) Risk assessment for invasive species produces net bioeconomic benefits. *Proceedings of the National Academy of Sciences*, **104**, 203–207.

Lobo, J. (2008) AUC : A misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, **17**, 145–151.

Mainali, K.P., Warren, D.L., Dhileepan, K., Mcconnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B.B., & Parmesan, C. (2015) Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. *Global Change Biology*, **21**, 4464–4480.

Le Maitre, D.C., Thuiller, W., & Schonegevel, L. (2008) Developing an approach to defining the potential distributions of invasive plant species: A case study of *Hakea* species in South Africa. *Global*

Ecology and Biogeography, **17**, 569–584.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43 SRC-**, 405–412.

Peterson, A.T. & Robins, C.R. (2003) Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology*, **17**, 1161–1165.

Phillips, S.J. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Phillips, S.J., Dudík, M., Dudik, M., & Phillips, S.J. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Source: Ecography*, **31**, 161–175.

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., & Maiorano, L. (2017) Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, **40**, 1076–1087.

Roy, H.E., Peyton, J., Aldridge, D.C., et al. (2014) Horizon scanning for invasive alien species with the potential to threaten biodiversity in Great Britain. *Global Change Biology*, **20**, 3859–3871.

Senay, S.D., Worner, S.P., & Ikeda, T. (2013) Novel three-step pesudo-abscence selection technique for improved species distribution modelling. *PLoS One*, **8**, e71218.

Siefert, A., Ravenscroft, C., Althoff, D., Alvarez-Yépiz, J.C., Carter, B.E., Glennon, K.L., Heberling, J.M., Jo, I.S., Pontes, A., Sauer, A., Willis, A., & Fridley, J.D. (2012) Scale dependence of vegetation-environment relationships: A meta-analysis of multivariate data. *Journal of Vegetation Science*, **23**, 942–951.

Storkey, J., Stratonovitch, P., Chapman, D.S., Vidotto, F., & Semenov, M.A. (2014) A process-based approach to predicting the effect of climate change on the distribution of an invasive allergenic plant in Europe. *PLoS ONE*, **9**, .

Tanner, R., Branquart, E., Brundu, G., Buholzer, S., Chapman, D., Ehret, P., Fried, G., Starfinger, U., & van Valkenburg, J. (2017) The prioritisation of a short list of alien plants for risk analysis within the framework of the Regulation (EU) No. 1143/2014. *NeoBiota*, **35**, 87–118.

Thuiller, W., Brotons, L., Araújo, M.B., & Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **27**, 165–172.

Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016) biomod2: Ensemble platform for species distribution modeling. R package version 3.3-7. *Available at: https://cran.r-project.org/web/packages/biomod2/index.html*, .

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

Václavík, T. & Meentemeyer, R.K. (2009) Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, **220**, 3248–3258.

VanDerWal, J., Shoo, L.P., Graham, C., & Williams, S.E. (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, **220**, 589–594.

Vilà, M., Espinar, J.L., Hejda, M., Hulme, P.E., Jarošík, V., Maron, J.L., Pergl, J., Schaffner, U., Sun, Y., & Pyšek, P. (2011) Ecological impacts of invasive alien plants: A meta-analysis of their effects on species, communities and ecosystems. *Ecology Letters*, **14**, 702–708.

Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. *Annals of Applied Statistics*, **4**, 1383–1402.

Zhao, X., Wang, G., Shen, Z., Zhang, H., & Qiu, M. (2006) Impact of elevated CO2 concentration under three soil water levels on growth of Cinnamomum camphora. *Journal of Zhejiang University,* *Science B*, **7**, 283–290.

Zomer, R.J., Trabucco, A., Bossio, D.A., & Verchot, L. V (2008) Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agr Ecosyst Environ*, **126**, 67–80.

Biosketch

The research team focuses on risk assessment for emerging invasive non-native species in Europe. Among other factors contributing to risk, the team use global-scale species distribution modelling to identify the suitable conditions for establishment by the focal species and use this to project their potential distributional range in the risk assessment area.