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, 2], Oliver L. Pescott[3], Helen E. Roy[3], Rob Tanner[4]

Institutional affiliations:

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

2 Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, UK

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

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

Corresponding author: Daniel Chapman

Email addresses: Daniel Chapman daniel.chapman@stir.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. We compared models where the pseudo-absences were selected from the accessible background, the unsuitable background (defined using biological knowledge of the species' key limiting factors) or from both types of background.

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, generally, 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 concern 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) necessitating 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; Elith et al., 2010; 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 (Elith et al., 2010; 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 well as prior biological knowledge about niche responses to environmental factors. The approach is based on combining 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 as only the accessible region, only the unsuitable region, 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 and randomly partitioned into training and testing datasets comprising 80% and 20% of the grid cells, respectively. 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 domains

We obtained background samples from both the accessible region and from the unsuitable region outside of the accessible region for each species (Figures 1-2). The effect was therefore to exclude potentially suitable but inaccessible regions from the combined background sample. For each of the five species in this study, ten replicate background samples were generated in order to reduce sampling variation (Barbet-Massin et al., 2012). 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. For evaluating the models by cross-validation, a randomly selected 20% of the accessible area for each species was added to its testing dataset and reserved from background sampling for model fitting. From the remaining accessible area, we drew the same number of background samples as there were occurrences (Barbet-Massin et al., 2012), weighted by the vascular plant record density as a target group. This ensured that the accessible area background sample used for model fitting 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 is not a relevant consideration in 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 unsuitable 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 model fitting, 3000 random samples were taken from the unsuitable region. If the unsuitable region overlapped with the accessible region, accessible parts of the unsuitable region were excluded. 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, only the unsuitable area or using the combined background samples from both the accessible and unsuitable areas. In all cases, model performance was evaluated by cross validation, using model predictions for 20% of the occurrences, accessible area and unsuitable area that were reserved from model fitting (the testing dataset).

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 within the training data (80:20% random split). 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, 2016). 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 representing the 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 or the unsuitable background were compared with those based on the combined accessible and unsuitable background in a standardised cross validation. To do this we used calculated the AUC for model projections on the 20% of occurrences versus the 20% of the accessible background that was reserved from model fitting and versus 20% of the accessible and unsuitable background. This allows comparison of projection accuracy within the range of the species and globally. As mentioned above, AUC in this context is informative about the relative discrimination power of different model specifications on the same data.

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). In most cases, cross-validated AUC indicated that models trained using samples from the combined accessible and unsuitable background were more accurate than those trained using only the individual accessible or unsuitable backgrounds (Table 1 and see Appendix S3). This was most clearly seen for global projections of the model, where the combined models had the highest AUC values for all five species (Table 1). The probability of the combined background model having the highest AUC of the three model types for all five species by chance is $P = 0.004$. For projections within the accessible range of the species, models sampling the combined accessible and unsuitable background were equally or marginally more accurate than models using only the accessible background in four out of five species, and always performed better than models using only the unsuitable background (Table 1).

Models using only the accessible background spanned a narrower range of suitability values and environmental conditions than the other two background specifications, and therefore their response curves were only constructed over a narrow range and provided a less clear distinction between high

and low suitability (Figure 3). Models using only the unsuitable background generated response curves that essentially discriminated unsuitable from suitable conditions completely, but did not capture suitability gradients within the range of the species (Figure 3). By contrast, the models based on the combined accessible and unsuitable background yielded partial response curves intermediate in form to the two individual background specifications (Figure 3), using information from the accessible background region to characterise gradients in suitability within the environmentally-suitable range of the species, and using the unsuitable background to identify conditions in which the species very rarely occurs. In some cases the models using combined accessible and unsuitable backgrounds yielded response curves that differed markedly from those of the accessible background models. This was most clearly seen in the responses of *Cinnamomum camphora* and *Lygodium japonicum* to low moisture (CMI), *Lespedeza cuneata* to low winter temperature (Bio6) and *Cinnamomum camphora* to high summer temperature (Bio10) (Figure 3).

Projections of potential non-native ranges from the models were strongly influenced by the choice of background specification (Figures 4 and 5, see Appendix S4 for global and native range projections). As was seen for the response curves, models based only on the accessible background generally made a gradual delineation between very low and high suitability, while models based on the unsuitable background made very sharp delineations and predicted larger invadable regions. Projections of models using the combined accessible and unsuitable backgrounds were intermediate in form, and represented gradients in suitability within the invaded regions as well as learning from the unsuitable background to rule out occurrence in those regions. For example, in North America the models using only the accessible background predicted suitability for *Cinnamomum camphora*, *Triadica sebifera* and *Lygodium japonicum* invasion in arid parts of south western USA. By contrast, models combining the accessible and unsuitable backgrounds suggested these regions were unsuitable for invasion (Figure 4). Similar effects could be seen in Europe, principally in terms of the effects of unsuitability rules about low winter temperature restricting suitability in central and Eastern Europe and rules about drought sensitivity restricting occurrence in Iberia (Figure 5). The projections also illustrated that models based

on only the accessible background were affected by extrapolation issues, resulting in gaps in the projected risk maps.

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; Barbet-Massin et al., 2012). 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 or just the unsuitable region.

This was evidenced by a consistent improvement in cross-validated discrimination power when the modelling sampled from a background combining accessible and biologically-informed unsuitable regions. This was most clearly seen in the global projections, where the combined background models always performed better than models using just the accessible or just the unsuitable background. For projections within the species' accessible range the combined background models gave consistently more accurate projections than models based only on the unsuitable background, and generally

performed as well as or marginally better than models trained only on the accessible background. Our expectation was that the combined background modelling strategy would not improve discrimination within the range of a species over models trained on the accessible region. 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 marginally improved performance for both purposes because we explicitly tried to exclude 'suitable-but-not-reached' locations from the larger background region by restricting it to locations considered environmentally unsuitable. 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. Further testing is required to determine whether a similar strategy would also benefit models for native as well as non-native species distribution models, but in principle our new strategy should confer similar advantages.

The influence of the accessible and unsuitable backgrounds on species-environment relationships was clearly seen in the response curves and projections of the different models. The combination of unsuitable and accessible backgrounds had four clear effects, when compared to the models using only the accessible background. 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 spanned a much wider range of environmental conditions than were found in the accessible background, which has previously been shown to be important for accurate spatial and temporal transfer of species distribution models (Guevara et al., 2017). Sampling unsuitable conditions only from within the accessible part of the species range would therefore require a greater amount of model extrapolation than our strategy does. Third, the response curves were less complex or multi-modal than those from models using only the accessible background (see responses for high CMI), which is more consistent with niche theory (Austin, 2002). Fourth, the response curves generally reflected prior assumptions about environmental limitation of the species and as such were more consistent with ecological understanding of the species.

For instance, combined background 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. These responses were not estimated by the model based only on the accessible background, but are 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, combining the accessible and unsuitable backgrounds led to models that 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 only case where the response curves did not always follow the rules defining unsuitability was for limitation by extremely high summer temperature. This may be because of a correlation between high summer and winter temperatures, the latter being limiting when high summer temperature was not. This suggests our approach may have sensitivity to collinearity in model predictors that requires further investigation (Dormann et al., 2012).

Nevertheless, the broader conclusion is that sampling from an unsuitable background, in addition to an accessible 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, our approach offers a simple 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). While there are more sophisticated approaches available for doing this using Bayesian models in which prior estimates of niche parameters can be specified (Talluto et al., 2015), a major advantage of the approach developed here is that it is implemented by manipulating the input data to standard distribution model software such as Biomod (Thuiller et al., 2009) or Maxent (Phillips et al., 2008) and all regression and machine learning methods. As such it is simple to implement with techniques that most modellers are

already familiar with and can quickly be applied in a standard way across species. This is especially useful when risk assessments are being performed across large numbers of invasive non-native species and require consistent judgements about establishment risk (Branquart et al., 2016; Tanner et al., 2017).

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). We recommend that similar sensitivity analyses are performed when applying our approach to other species, since previous studies have found these factors can strongly influence distribution model performance (Barve et al., 2011; Barbet-Massin et al., 2012). 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. 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. Cross-validated discrimination performance of ensemble model projections for the potential global distribution of five plant species developed using different background region specifications (A = accessible background, U = unsuitable background, A&U = combined accessible and unsuitable background). Discrimination performance is the cross-validated AUC (Area Under the receiver-operator Curve) and its standard deviation in parentheses for model predictions on 20% of the occurrences, accessible background and unsuitable background that were reserved from model fitting. For presence-only data AUC is the probability that a species presence has a higher projected suitability than a background sample.

Species	**Accuracy in the species range (AUC in accessible background)**			**Global accuracy (AUC in the accessible and unsuitable backgrounds)**		
	A	**U**	**A&U**	**A**	**U**	**A&U**
Cinnamomum camphora	0.664 (0.019)	0.581 (0.020)	0.669 (0.020)	0.857 (0.008)	0.981 (0.001)	0.985 (0.001)
Humulus scandens	0.742 (0.020)	0.669 (0.017)	0.737 (0.021)	0.977 (0.004)	0.979 (0.001)	0.982 (0.002)
Lespedeza cuneata	0.899 (0.006)	0.860 (0.006)	0.899 (0.006)	0.979 (0.003)	0.977 (0.002)	0.983 (0.002)
Lygodium japonicum	0.852 (0.013)	0.758 (0.014)	0.852 (0.013)	0.955 (0.007)	0.979 (0.001)	0.987 (0.001)
Triadica sebifera	0.762 (0.017)	0.673 (0.016)	0.777 (0.017)	0.853 (0.007)	0.984 (0.001)	0.989 (0.001)

Figures

Figure 1. Parts of the regions from which background samples (pseudo-absences) were drawn for modelling one of the five species, *Humulus scandens*. Shading shows the accessible background, where the species is assumed to have had chance to disperse to and sample, and the unsuitable background, defined using biological information on the key limiting factors of the species (see Appendix S2). Potentially suitable, but inaccessible locations were excluded from the modelling (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 highly restricted to represent stronger dispersal constraints during the invasive range expansion.

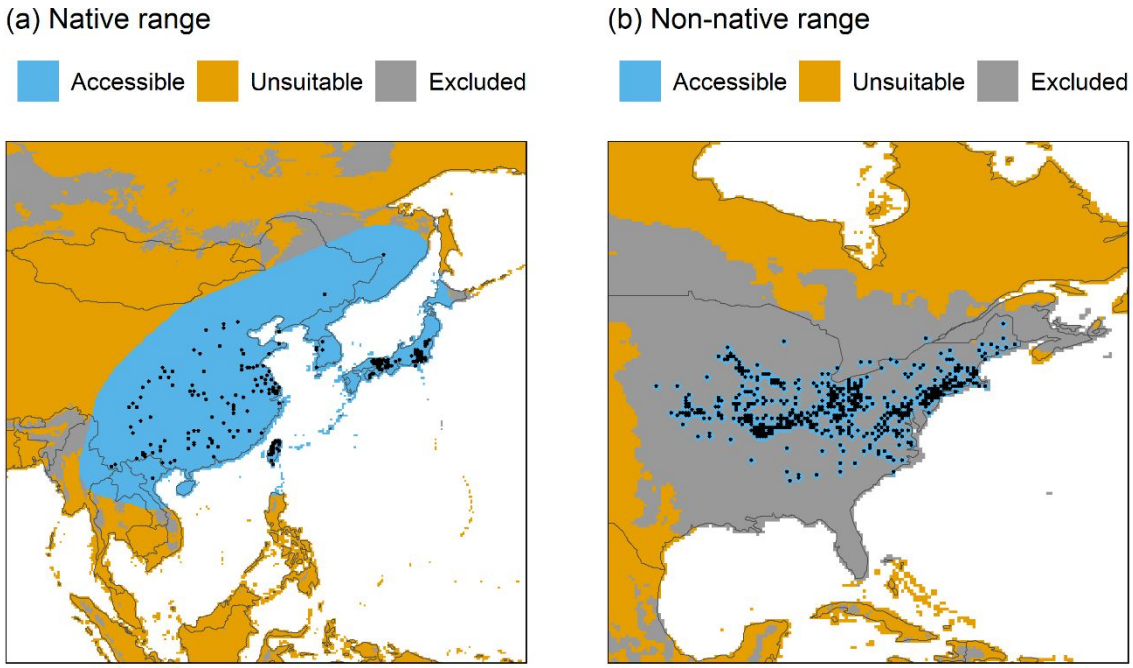


Figure 2. Flow chart for implementing the biologically-informed pseudo-absence selection for presence-background modelling of invasive non-native species.

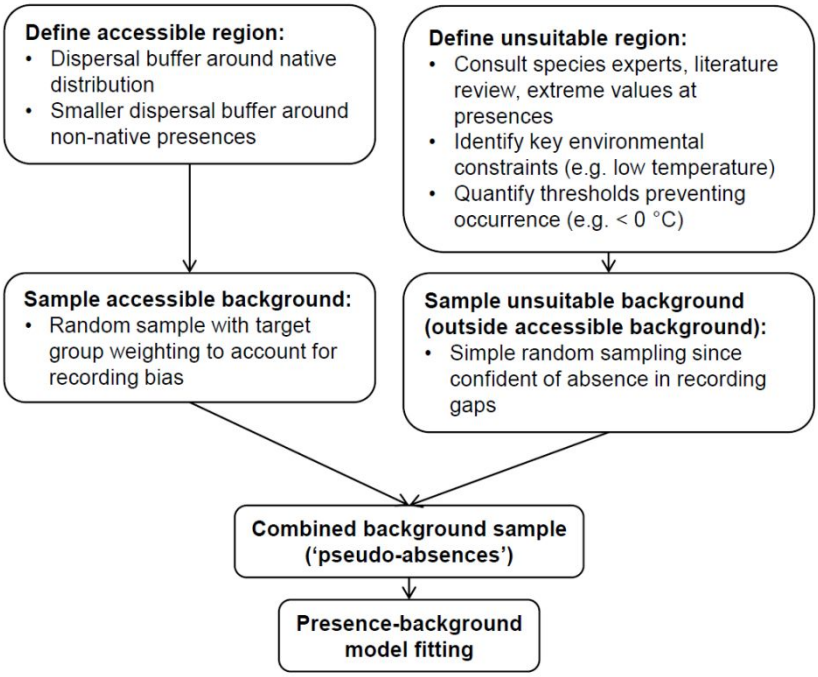


Figure 3. Partial response plots fitted by the ensemble models showing the predicted suitability when other variables are fixed at suitable values for the species (medians in the presence grid cells). Curves span the range of the variables in the training data. Curve colour differentiates the models 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).

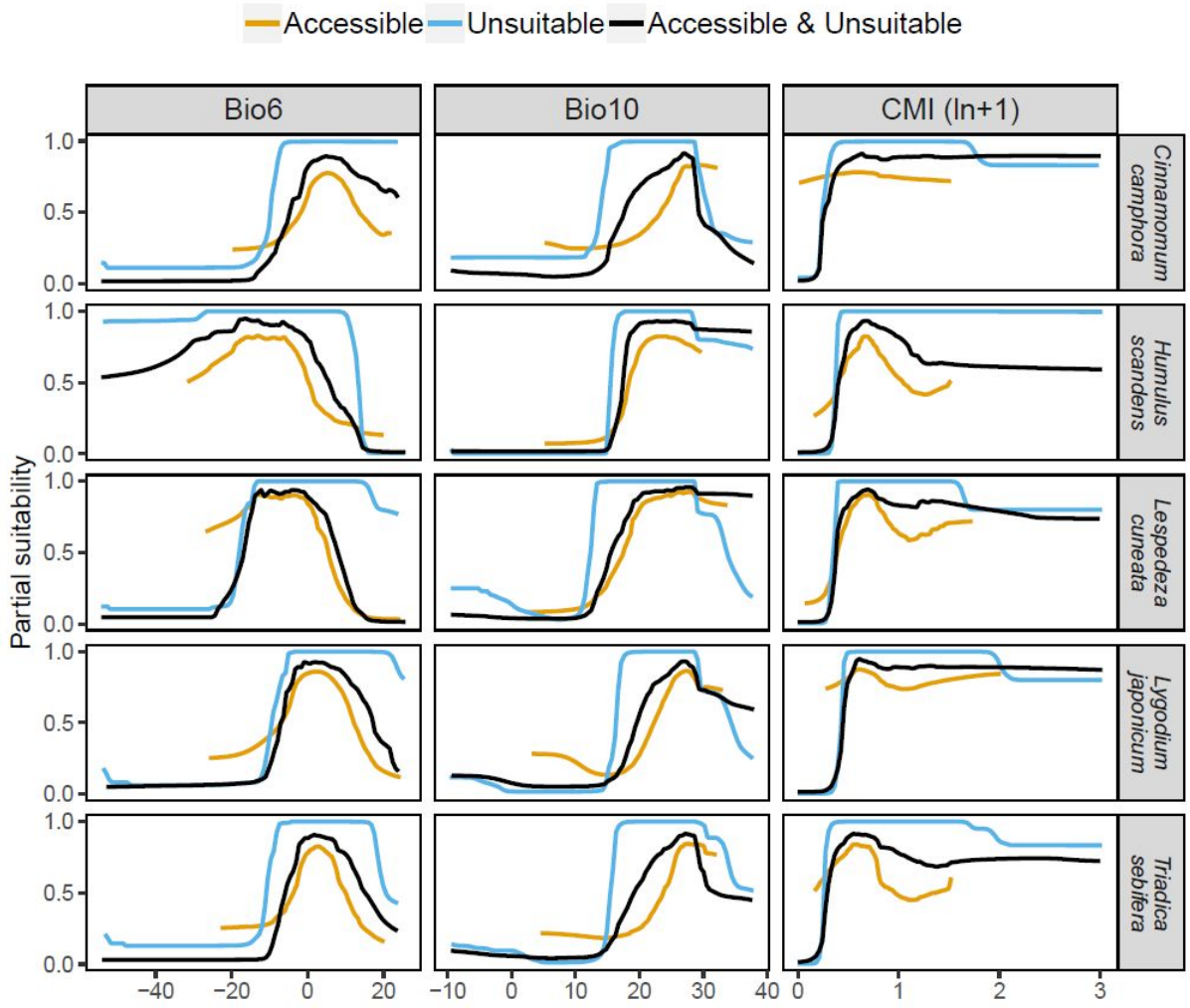


Figure 4. Potential non-native distributions of five Asian plant species in the USA, where all are already established invasive non-native species with expanding ranges. Projections are from models where the background domain is either just the accessible area, just the unsuitable area or the combined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

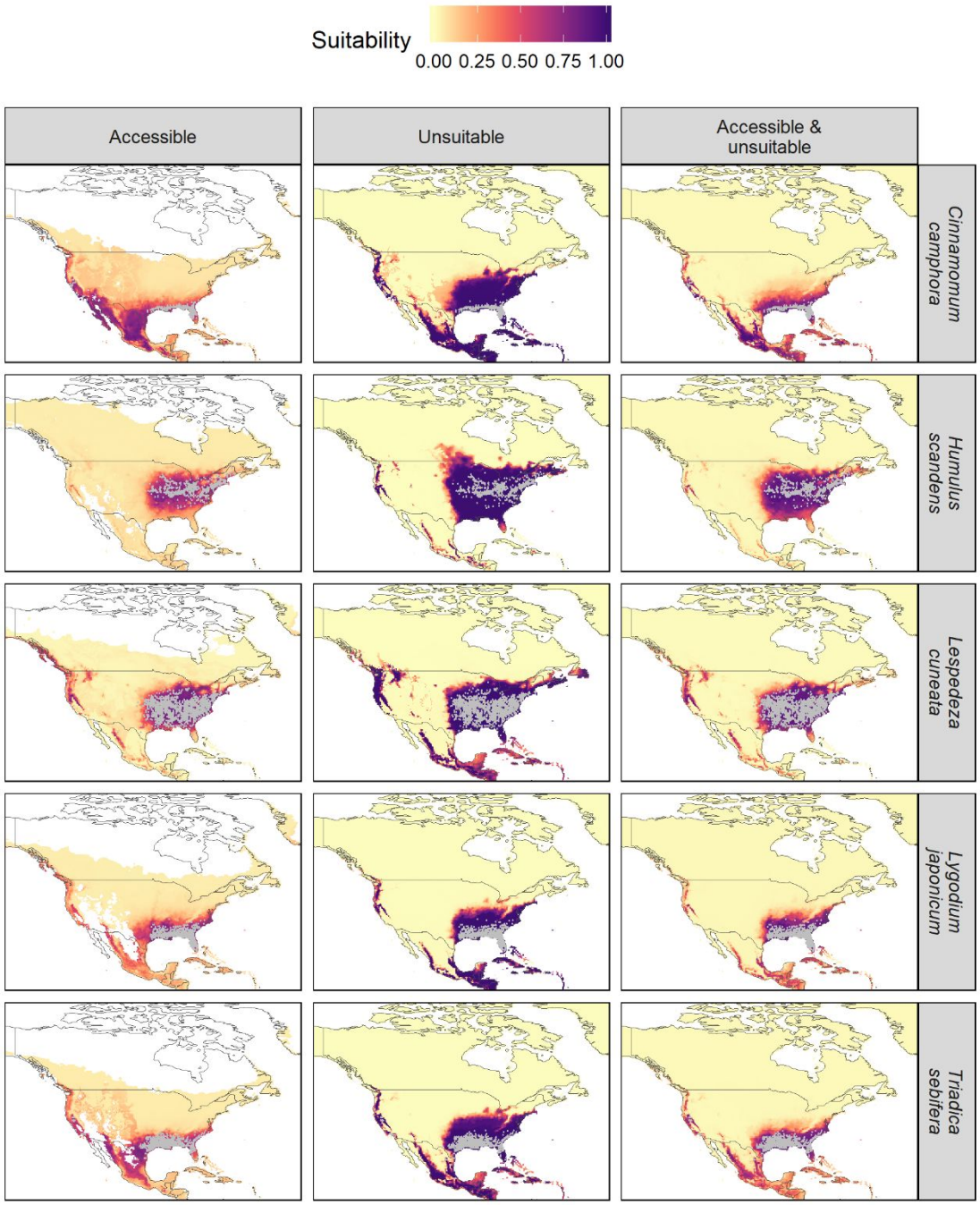
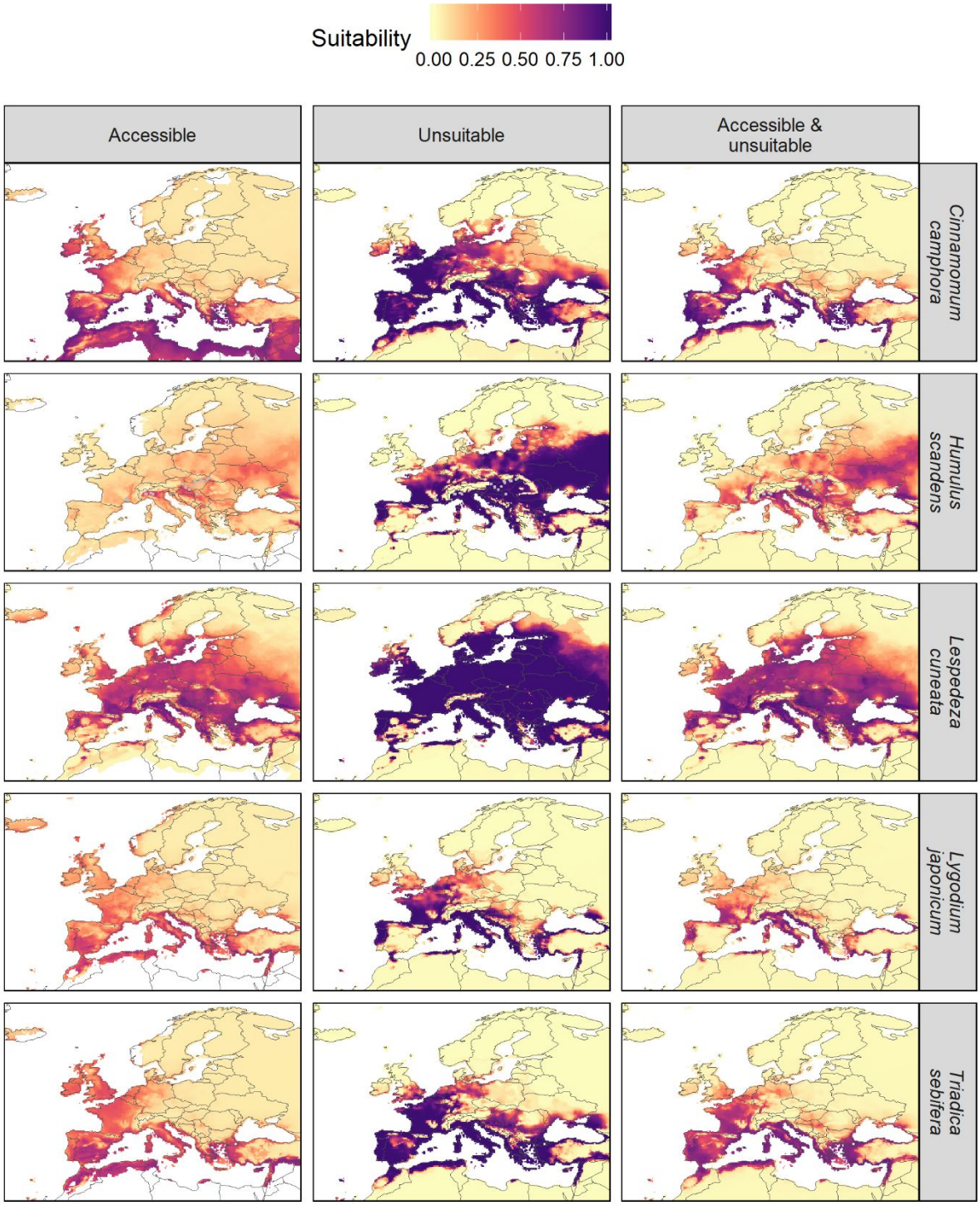


Figure 5. Potential distributions of five Asian plant species in Europe, where the species are currently absent or emerging invasive non-native species, equivalent to Figure 4.



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.

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.

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., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., & Lautenbach, S. (2012)

Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

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.

Elith, J., Kearney, M., & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

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/.

Guevara, L., Gerstner, B.E., Kass, J.M., & Anderson, R.P. (2017) Toward ecologically realistic predictions of species distributions: A cross-time example from tropical

montane cloud forests. *Global Change Biology*, **24**, 1511–1522.

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. *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**, e88156.

Talluto, M. V, Boulangeat, I., Ameztegui, A., Aubin, I., Berteaux, D., Butler, A., Doyon, F., Drever, C.R., Fortin, M.-J., Franceschini, T., Liénard, J., McKenney, D., Solarik, K.A., Strigul, N., Thuiller, W., & Gravel, D. (2015) Cross-scale integration of knowledge for predicting species ranges: a metamodelling framework. *Global Ecology and Biogeography*, **25**, 238–249.

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.

Data Accessibility

A data file containing the 0.25 x 0.25 gridded data on climate, recording effort, species occurrence, accessibility and unsuitability is included in the Supporting Information.

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.

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

Daniel Chapman, Oliver L. Pescott, Helen E. Roy, Rob Tanner

Appendix S1 – Proxy for recording effort

Figure S1.1. The global density of vascular plant (phylum Tracheophyta) records retrieved from the Global Biodiversity Information Facility, mapped on a 1x1 degree grid and displayed on a $\log_{10}$ scale. Dark grey areas returned no records.

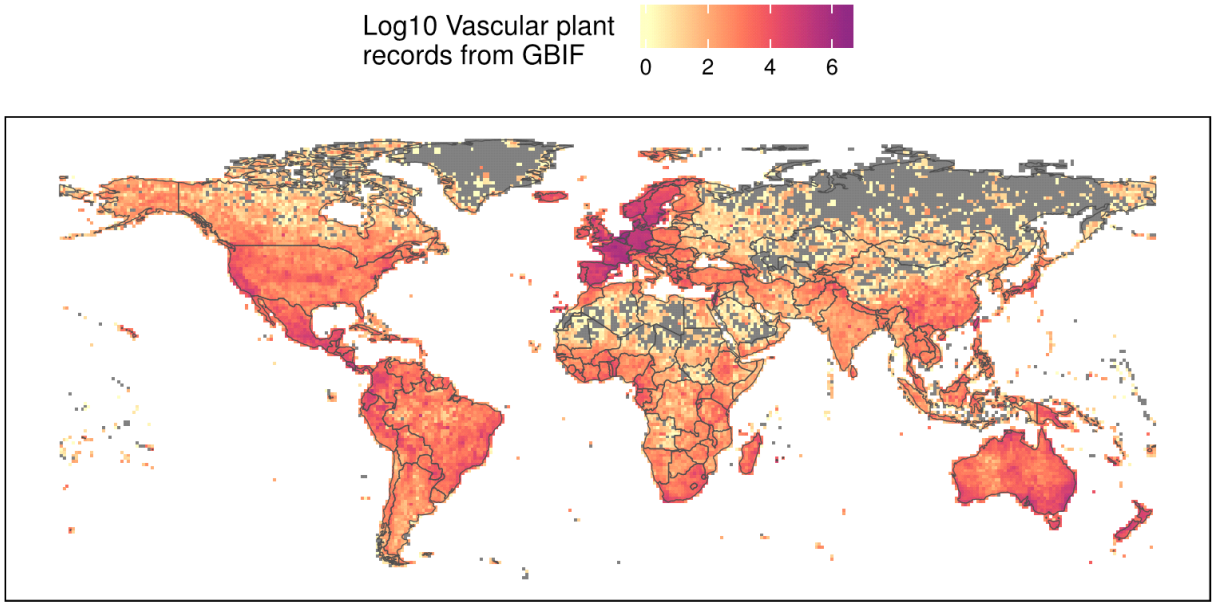
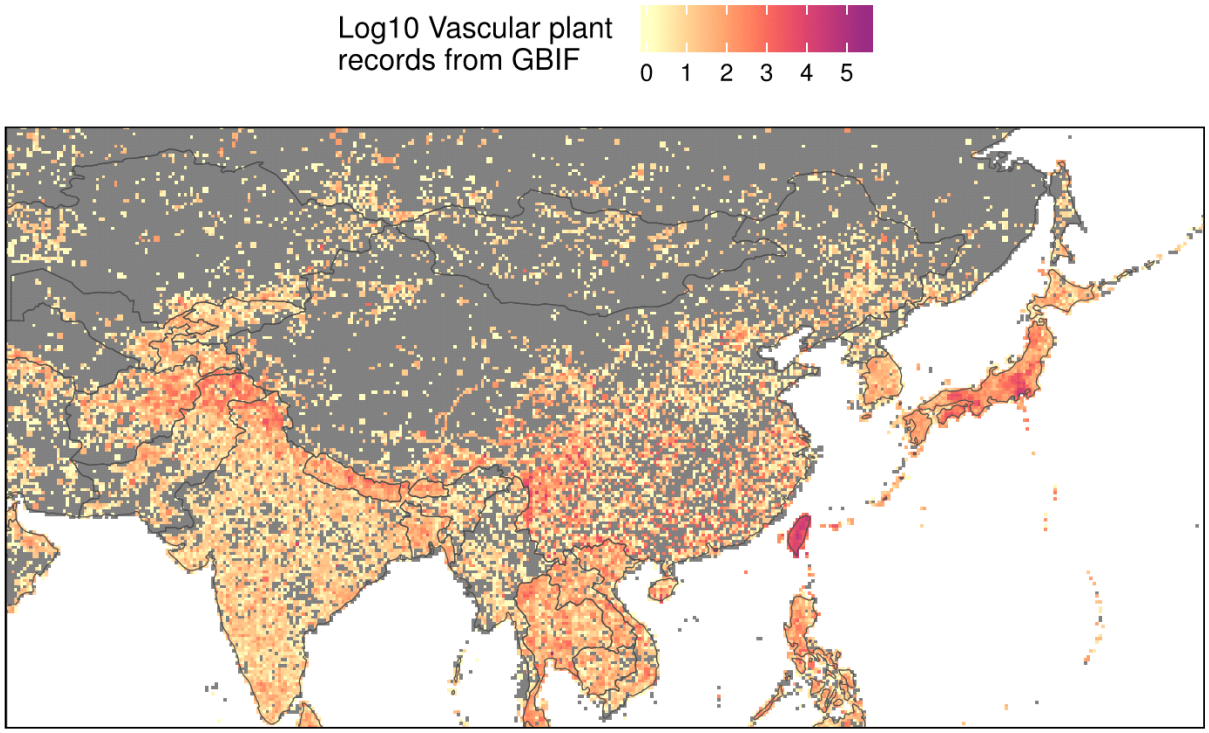


Figure S1.2. The density of vascular plant (phylum Tracheophyta) records retrieved from the Global Biodiversity Information Facility at the spatial resolution of the model (0.25 x 0.25 degree grid) for Asia. This region spanned the native range of most of the study species and had pronounced recording bias. Dark grey areas returned no records.



Appendix S2 – Distributions of the five modelled plant species and definitions of environmentally-unsuitable regions

Figure S2.3. Distribution records of the five species that were modelled, plotted as black points. Background shading shows the accessible and unsuitable background domains from which 'pseudo-absences' were drawn. The accessible background is only visible in the native range of the species, as it is otherwise masked by the non-native presences. Our approach excludes parts of the world that may be environmentally suitable, but are out of dispersal range (unshaded).

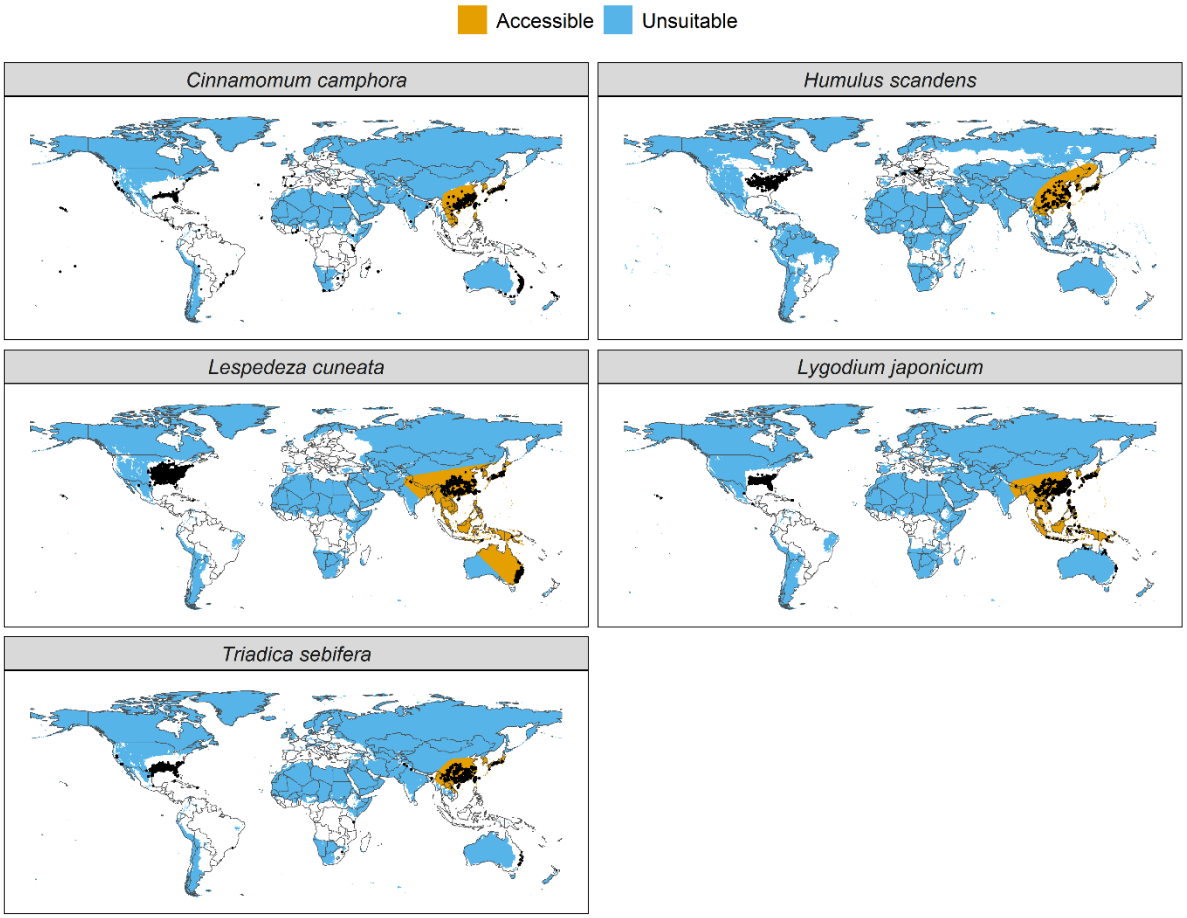


Table S2.1. Rules for defining highly unsuitable conditions for establishment by the five study species, with respect to three climatic variables. The rules are based on a combination of prior biological knowledge about key constraints and the extreme values of the climate variables at the species occurrence locations. For CMI and upper limit on Bio10, few direct estimates were used, so the 0.5^{th} or 99.5^{th} percentile values at occurrence locations were used as thresholds, respectively. The individual rules were combined with an OR statement to generate unsuitable background domains for the distribution modelling.

Species	**Bio6 (mean minimum temperature during coldest month)**	**Bio10 (mean temperature during warmest quarter)**	**CMI (climatic moisture index)**
Cinnamomum camphora	<-10 °C; the temperature causing frost damage in overwintering seedlings (You et al., 2008).	<15 °C; the reported minimum annual temperature (Orwa et al., 2009; CABI, 2018), **OR** >29 °C	<0.25
Humulus scandens	>16 °C; overwintering seeds require stratification (Balogh & Dancza, 2008) and the warmest occurrence is at 15.5 °C.	<15 °C; approximately corresponds to a known requirement of ~1300 degree days (base 4 °C) for maturation (G. Fried, personal communication), **OR** >28 °C	<0.45
Lespedeza cuneata	<-17 °C; effects of frost are uncertain but may contribute to mortality (Gucker, 2010) and the coldest occurrence is at 16.4 °C.	<13 °C; consistent with minimum temperatures for germination (Qiu et al., 1995) and seedling growth (Hill & Luck, 1991) , **OR** >29 °C; consistent with observed reductions in leaf size and height when grown above 25 °C (Kalburtji et al., 2007)	<0.45
Lygodium japonicum	<-8 °C; consistent with mean values of Bio6 in USDA Plant Hardiness Zone 6, where the species is semi-hardy (Loan, 2006).	<16 °C; no information available on summer temperature requirements, so the 0.5^{th} percentile value of occurrences was assumed to be a low limit, **OR** >29 °C	<0.55
Triadica sebifera	<-9 °C; frost is considered the strongest factor limiting invasion in the USA (Gan et al., 2009) and the coldest occurrence is at -8.5 °C.	< 16 °C; consistent with the temperature inhibiting seed germination (Nijjer et al., 2002) , **OR** >29 °C	<0.30

Appendix S3 – Model summary tables

Table S3.2. Details of the ensemble presence-background models for five Asian plant species occurring elsewhere as invasive non-native species. Three model specifications were applied, where the background sample was drawn from only an accessible background, only from an unsuitable background or from a combined accessible and unsuitable background. Seven individual algorithms were fitted and evaluated based on their AUC, cross-validated within the training data. Poorly performing algorithms were rejected, and the remaining ones combined into the ensemble. Variable importances are given as percentages for Bio6 (mean minimum temperature of coldest month), Bio10 (mean temperature of warmest quarter) and CMI (ratio of annual precipitation to potential evapotranspiration).

		Accessible background					**Unsuitable background**					**Accessible and unsuitable background**				
Species	**Algorithm**	**AUC**	**In ensemble**	**Bio6**	**Bio10**	**CMI**	**AUC**	**In ensemble**	**Bio6**	**Bio10**	**CMI**	**AUC**	**In ensemble**	**Bio6**	**Bio10**	**CMI**
Cinnamomum camphora	ANN	0.758	yes	74%	12%	13%	0.998	yes	39%	31%	30%	0.946	yes	50%	24%	26%
	GAM	0.753	yes	93%	4%	3%	0.997	yes	43%	29%	28%	0.941	no	45%	30%	24%
	GBM	0.771	yes	83%	8%	8%	0.992	no	42%	33%	25%	0.952	yes	48%	32%	21%
	GLM	0.750	yes	89%	6%	4%	0.996	yes	43%	29%	28%	0.937	no	51%	26%	24%
	MARS	0.750	yes	91%	3%	6%	0.996	yes	44%	28%	28%	0.948	yes	49%	28%	23%
	Maxent	0.749	yes	77%	9%	14%	0.998	yes	28%	44%	27%	0.952	yes	49%	31%	21%
	RF	0.726	no	52%	23%	24%	0.999	yes	36%	34%	30%	0.948	yes	48%	27%	25%
	Ensemble	**0.769**		**85%**	**7%**	**8%**	**0.999**		**39%**	**33%**	**29%**	**0.954**		**49%**	**28%**	**23%**
Humulus scandens	ANN	0.773	yes	55%	20%	25%	0.999	yes	24%	41%	34%	0.955	yes	34%	33%	33%
	GAM	0.773	yes	57%	19%	24%	0.998	yes	23%	44%	33%	0.957	yes	34%	33%	33%
	GBM	0.785	yes	54%	16%	30%	0.999	yes	22%	48%	30%	0.958	yes	37%	27%	36%
	GLM	0.765	yes	46%	26%	28%	0.999	yes	27%	40%	33%	0.953	yes	32%	35%	33%
	MARS	0.771	yes	56%	18%	27%	1.000	yes	22%	45%	33%	0.956	yes	34%	31%	35%
	Maxent	0.762	yes	54%	20%	26%	0.995	no	18%	48%	34%	0.955	yes	40%	27%	32%
	RF	0.762	yes	42%	26%	32%	0.998	yes	22%	44%	34%	0.953	yes	38%	28%	34%
	Ensemble	**0.784**		**52%**	**21%**	**27%**	**1.000**		**23%**	**44%**	**33%**	**0.959**		**36%**	**30%**	**34%**

		Accessible background					Unsuitable background					Accessible and unsuitable background				
Species	**Algorithm**	**AUC**	**In ensemble**	**Bio6**	**Bio10**	**CMI**	**AUC**	**In ensemble**	**Bio6**	**Bio10**	**CMI**	**AUC**	**In ensemble**	**Bio6**	**Bio10**	**CMI**
Lespedeza cuneata	ANN	0.912	yes	49%	11%	40%	0.999	yes	33%	30%	37%	0.963	yes	40%	21%	39%
	GAM	0.910	yes	47%	10%	43%	0.999	yes	37%	24%	39%	0.961	yes	41%	20%	38%
	GBM	0.914	yes	59%	6%	34%	0.999	yes	37%	30%	34%	0.964	yes	43%	19%	38%
	GLM	0.893	no	53%	11%	36%	1.000	yes	36%	25%	39%	0.953	no	42%	20%	38%
	MARS	0.907	yes	48%	11%	41%	0.998	yes	36%	25%	39%	0.963	yes	39%	20%	41%
	Maxent	0.904	yes	40%	18%	42%	0.998	yes	18%	44%	38%	0.962	yes	39%	25%	36%
	RF	0.904	yes	41%	20%	39%	0.999	yes	24%	35%	41%	0.961	yes	36%	25%	38%
	Ensemble	**0.912**		**47%**	**13%**	**40%**	**1.000**		**32%**	**30%**	**38%**	**0.964**		**40%**	**22%**	**38%**
Lygodium japonicum	ANN	0.853	yes	75%	16%	9%	0.998	yes	27%	31%	42%	0.959	yes	41%	26%	33%
	GAM	0.844	yes	89%	9%	2%	0.999	yes	33%	24%	43%	0.958	yes	38%	23%	39%
	GBM	0.857	yes	74%	20%	5%	0.997	yes	34%	31%	35%	0.965	yes	39%	29%	32%
	GLM	0.844	yes	84%	15%	1%	0.999	yes	30%	26%	45%	0.949	no	36%	24%	40%
	MARS	0.850	yes	78%	19%	2%	0.997	yes	28%	28%	44%	0.964	yes	39%	22%	39%
	Maxent	0.841	yes	64%	24%	12%	0.995	no	10%	45%	45%	0.962	yes	39%	30%	31%
	RF	0.838	yes	55%	24%	21%	0.999	yes	20%	36%	44%	0.961	yes	34%	32%	34%
	Ensemble	**0.855**		**74%**	**18%**	**8%**	**0.999**		**29%**	**29%**	**42%**	**0.966**		**38%**	**27%**	**35%**
Triadica sebifera	ANN	0.754	yes	64%	16%	21%	0.999	yes	36%	30%	34%	0.940	yes	44%	24%	32%
	GAM	0.756	yes	78%	13%	9%	1.000	yes	42%	26%	32%	0.933	yes	44%	24%	32%
	GBM	0.768	yes	64%	17%	19%	0.998	yes	40%	30%	30%	0.940	yes	46%	29%	25%
	GLM	0.750	yes	78%	15%	7%	1.000	yes	37%	29%	34%	0.929	no	44%	23%	32%
	MARS	0.750	yes	76%	14%	10%	0.998	yes	39%	29%	31%	0.937	yes	50%	23%	27%
	Maxent	0.755	yes	69%	16%	15%	0.991	no	16%	50%	34%	0.937	yes	45%	31%	24%
	RF	0.719	no	45%	28%	27%	0.998	yes	26%	40%	35%	0.931	no	42%	28%	29%
	Ensemble	**0.766**		**71%**	**15%**	**14%**	**1.000**		**37%**	**31%**	**33%**	**0.941**		**46%**	**26%**	**28%**

Appendix S4 – Native range and global projections for five modelled plant species

Figure S4.4. Projections of suitability for the five study species in Asia, which includes their main native distributions. Plots show outputs from the three background specifications, equivalent to Figure 4 in the main text.

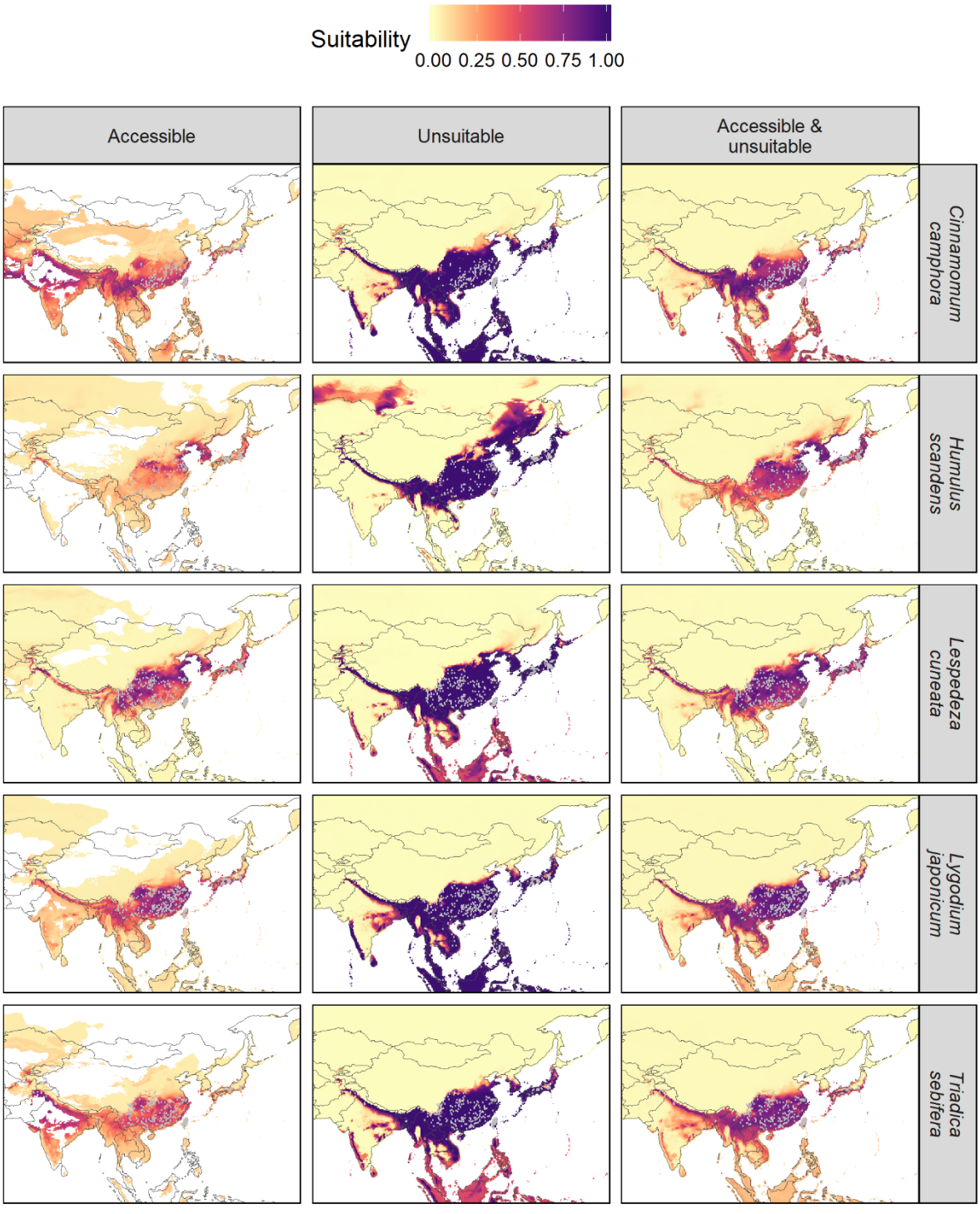
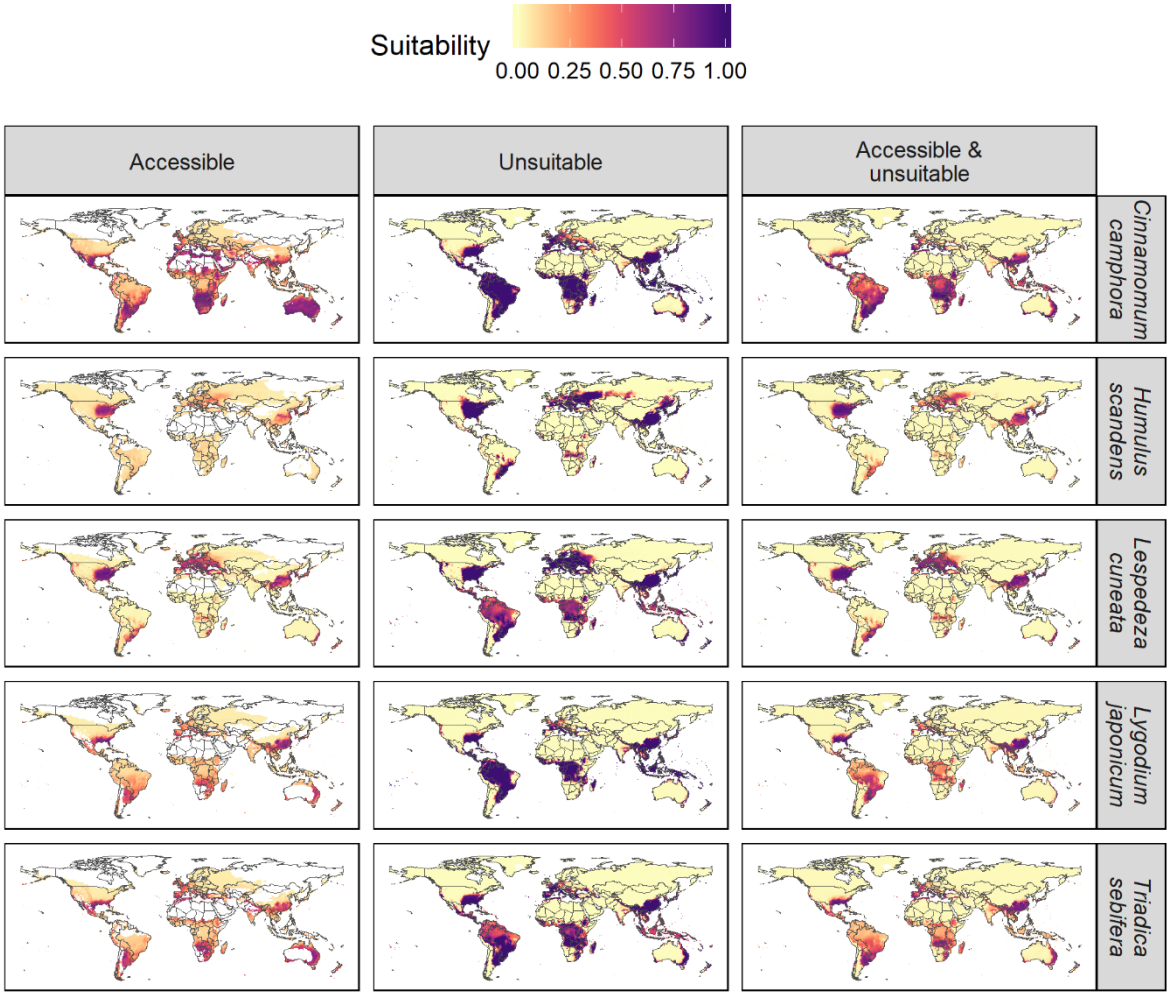


Figure S4.5. Global projections of suitability for the five study species. Plots show outputs from the three model specifications for the background region, equivalent to Figure 4 in the main text.



Appendix S5 – Sensitivity analysis on model settings

Sensitivity to size of the accessible region and number of background samples

All five species were modelled as in the main text but with all combinations of:

- the buffer radius for the native accessible region set to 0 km (unsuitable area model), 200 km, 400 km (as used in the main text) and 600 km; and
- the number of background samples (pseudo-absences) taken from the unsuitable region set to 0 (accessible area model), 1000, 3000 (as used in the main text) and 5000.

Model response plots were generally not very sensitive to the choice of these settings, except when only 1000 unsuitable background samples were taken (Figures S5.6-5.8). As a result, global suitability projections were almost identical (not shown).

Sensitivity to the rules used to define the unsuitable region

All five species were modelled as in the main text but with the rules defining their unsuitable regions set to either the values in Table S2.1 (used in the main text) or to a more conservative definition of the unsuitable region. For the latter, temperature thresholds were made more extreme by 2 °C (e.g. <-10 °C changed to <-12 °C; >16 °C changed to >18 °C, etc.) and moisture (CMI) thresholds were made more extreme by 10% (e.g. 0.25 changed to 0.15, etc.). The effect of this was to reduce the size of the unsuitable region and separate it more strongly from the species' occurrences. Surprisingly, this generally had little influence on the fitted response functions (Figure S5.9) or projections made from the models (not shown).

Figure S5.6. Partial response plots for models fitted using only accessible backgrounds, and with the native accessible region defined with buffer radii of 200 km, 400 km and 600 km.

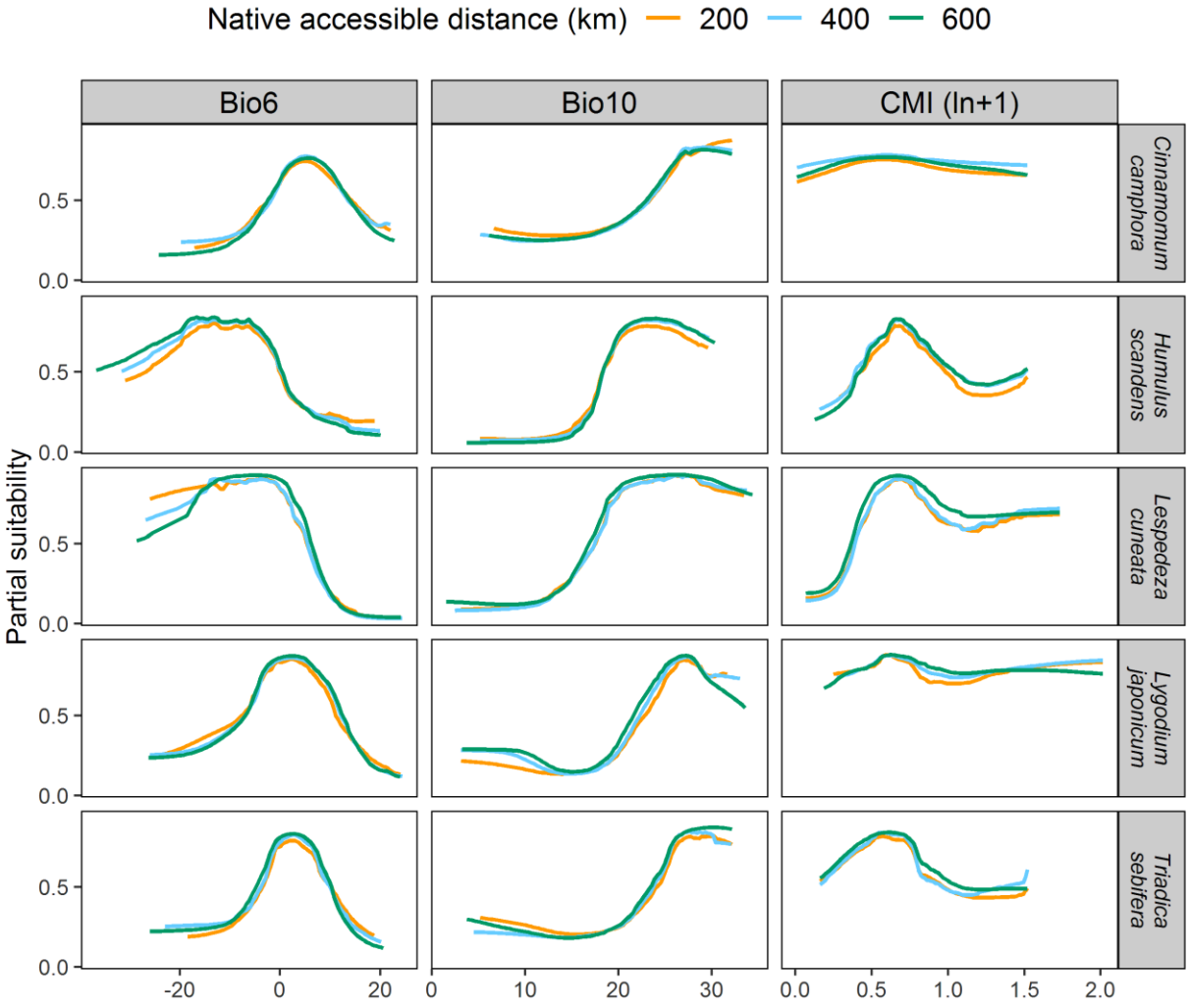


Figure S5.7. Partial response plots for models fitted using accessible and unsuitable backgrounds, and with the native accessible region defined with buffer radii of 200 km, 400 km and 600 km and 3000 background samples.

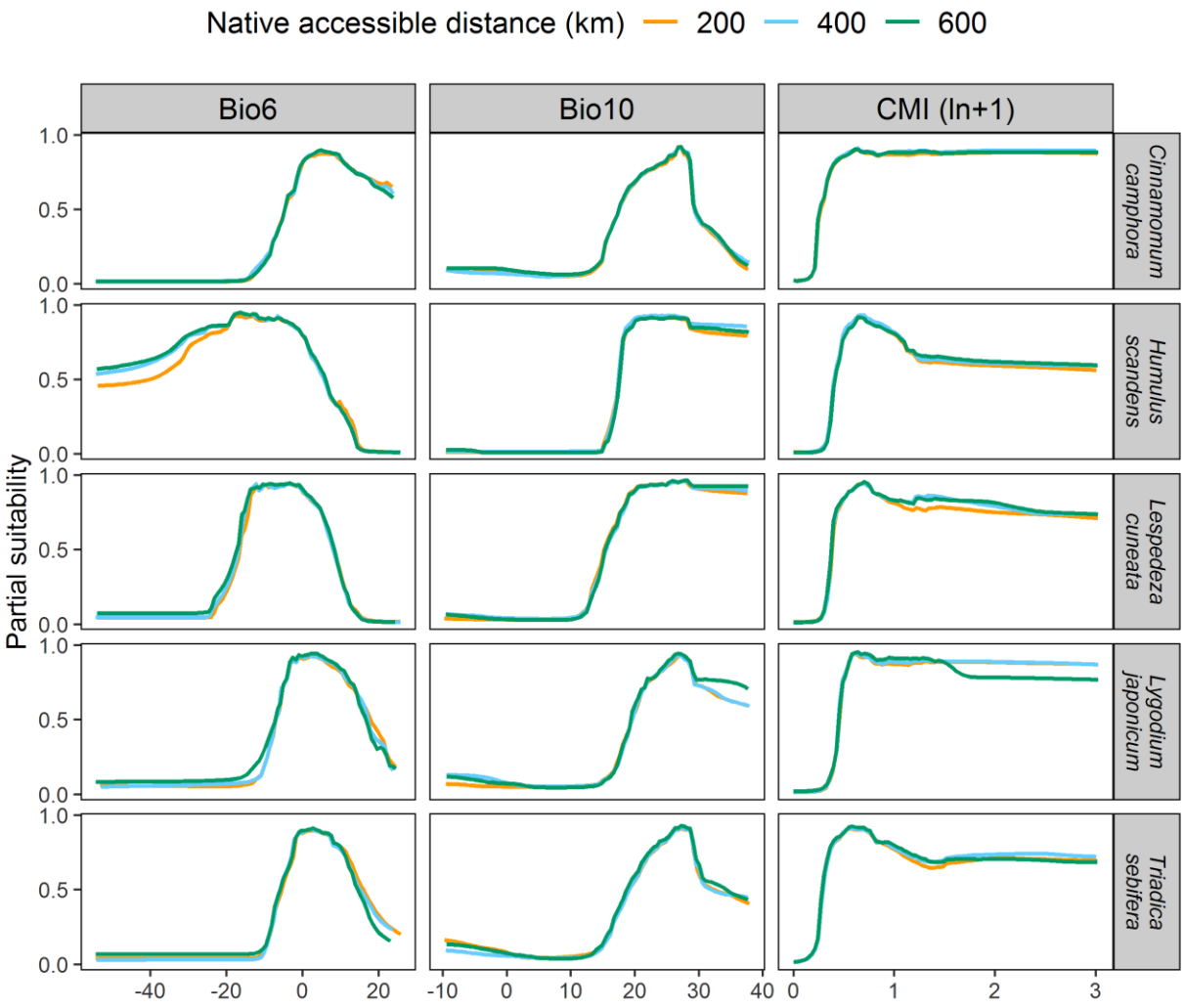


Figure S5.8. Effect of the number of unsuitable background samples on response plots fitted by models using accessible and unsuitable backgrounds. Models were fitted with 1000, 3000 or 5000 unsuitable background samples and a 400 km buffer for the native accessible region.

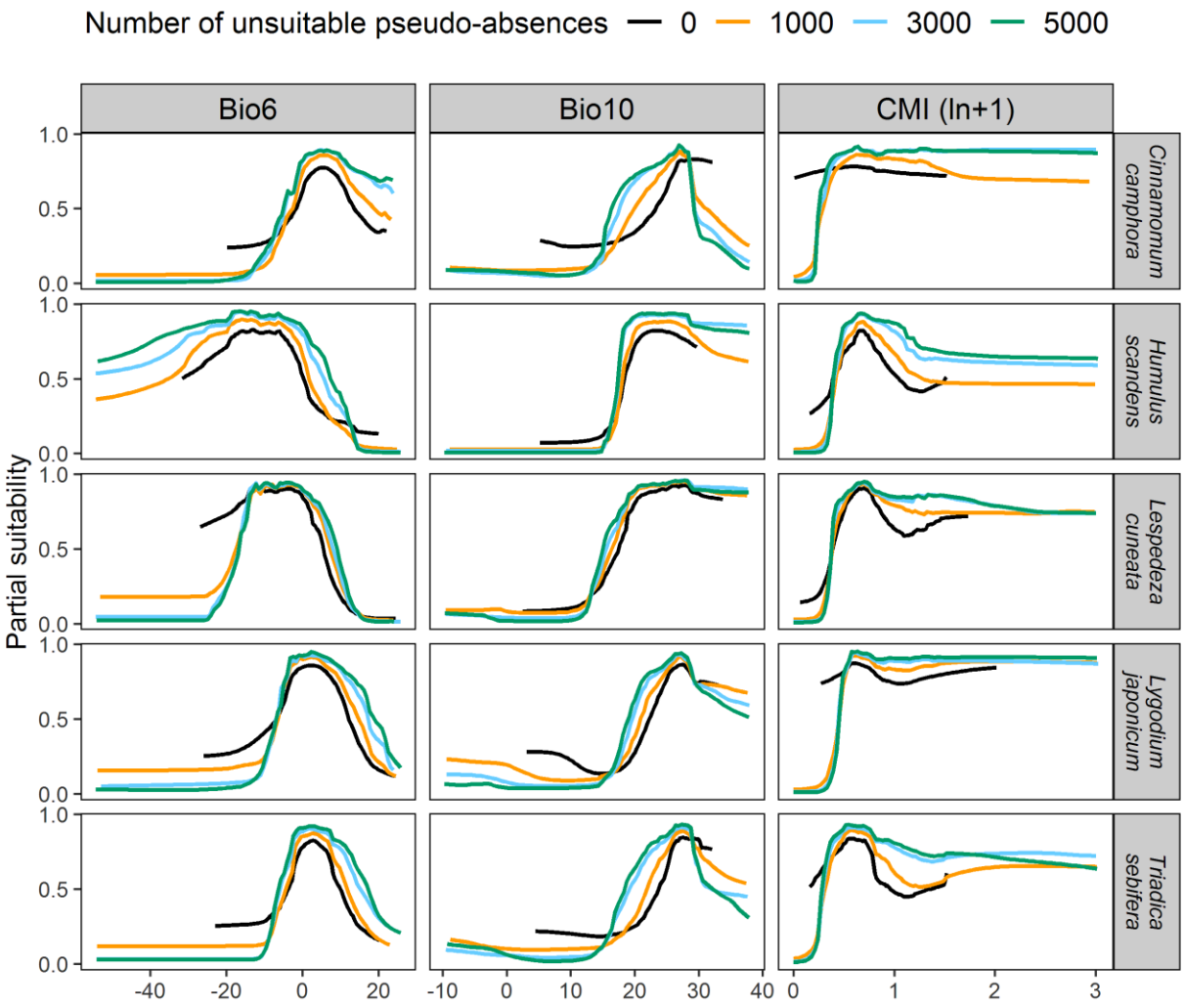
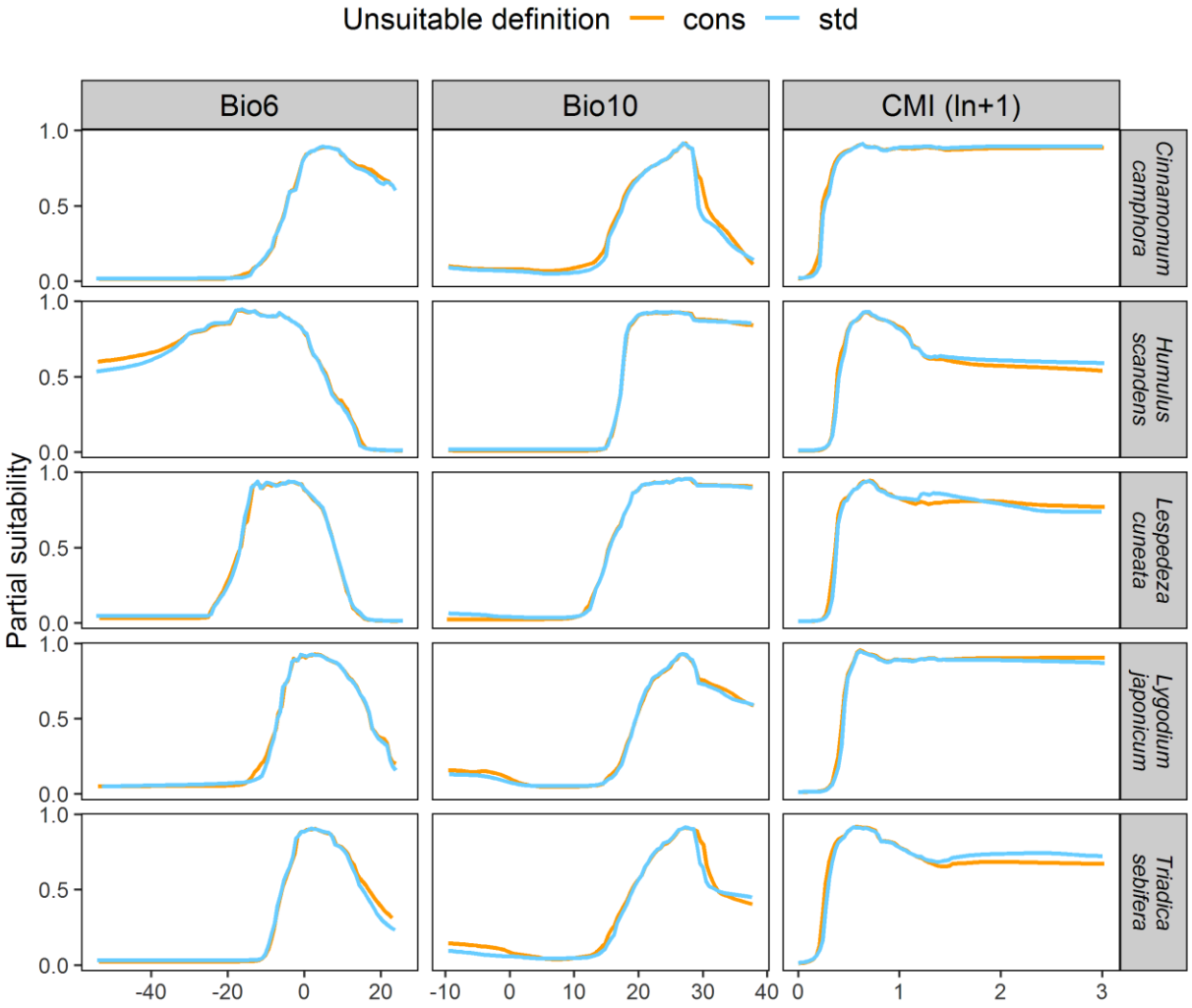


Figure S5.9. Response curves fitted to models where the unsuitable region was defined as in the main text (std) or with more conservative rules (cons).



Appendix S6 – 'Read me' for the data file

The file 'INNS SDM data for modelling.rds' contains the data used in the study, compiled from publicly available sources and gridded at 0.25 x 0.25 degree resolution. To open the file in R and read in the data as a data.frame please use the command:

```
readRDS('INNS SDM data for modelling.rds')
```

The .rds file contains a compressed R data.frame with the following columns:

- x = longitude of 0.25 degree grid cell centre
- y = latitude of 0.25 degree grid cell centre
- bio6 = Worldclim Bio6 (Minimum temperature of the coldest quarter, C)
- bio10 = Worldclim Bio10 (Mean temperature of the warmest quarter, C)
- moisture = climatic moisture index (ratio of annual precipitation to potential evapotranspiration)
- effort = proxy for recording effort, the number of Tracheophyte records held by GBIF
- occ.Cinnamomum.camphora = occurrences of *Cinnamomum camphora* (1 = occurrence)
- occ.Humulus.scandens = occurrences of *Humulus scandens* (1 = occurrence)
- occ.Lespedeza.cuneata = occurrences of *Lespedeza cuneata* (1 = occurrence)
- occ.Lygodium.japonicum = occurrences of *Lygodium japonicum* (1 = occurrence)
- occ.Triadica.sebifera = occurrences of *Triadica sebifera* (1 = occurrence)
- native.occ.Cinnamomum.camphora = occurrences of *Cinnamomum camphora* in native range (1 = occurrence)
- native.occ.Humulus.scandens = occurrences of *Humulus scandens* in native range (1 = occurrence)
- native.occ.Lespedeza.cuneata = occurrences of *Lespedeza cuneata* in native range (1 = occurrence)
- native.occ.Lygodium.japonicum = occurrences of *Lygodium japonicum* in native range (1 = occurrence)
- native.occ.Triadica.sebifera = occurrences of *Triadica sebifera* in native range (1 = occurrence)

- nonnative.occ.Cinnamomum.camphora = occurrences of *Cinnamomum camphora* in non-native range (1 = occurrence)
- nonnative.occ.Humulus.scandens = occurrences of *Humulus scandens* in non-native range (1 = occurrence)
- nonnative.occ.Lespedeza.cuneata = occurrences of *Lespedeza cuneata* in non-native range (1 = occurrence)
- nonnative.occ.Lygodium.japonicum = occurrences of *Lygodium japonicum* in non-native range (1 = occurrence)
- nonnative.occ.Triadica.sebifera = occurrences of *Triadica sebifera* in non-native range (1 = occurrence)
- accessible.Cinnamomum.camphora = the accessible background region for *Cinnamomum camphora* (1 = accessible)
- accessible.Humulus.scandens = the accessible background region for *Humulus scandens* (1 = accessible)
- accessible.Lespedeza.cuneata = the accessible background region for *Lespedeza cuneata* (1 = accessible)
- accessible.Lygodium.japonicum = the accessible background region for *Lygodium japonicum* (1 = accessible)
- accessible.Triadica.sebifera = the accessible background region for *Triadica sebifera* (1 = accessible)
- unsuitable.Cinnamomum.camphora = the unsuitable background region for *Cinnamomum camphora* (1 = unsuitable)
- unsuitable.Humulus.scandens = the unsuitable background region for *Humulus scandens* (1 = unsuitable)
- unsuitable.Lespedeza.cuneata = the unsuitable background region for *Lespedeza cuneata* (1 = unsuitable)

- unsuitable.Lygodium.japonicum = the unsuitable background region for *Lygodium japonicum* (1 = unsuitable)
- unsuitable.Triadica.sebifera = the unsuitable background region for *Triadica sebifera* (1 = unsuitable)

Supporting References

Balogh, L. & Dancza, I. (2008) *Humulus japonicus*, an emerging invader in Hungary. *Plant invasions: Human perception, ecological impacts and management* (ed. by B. Tokarska-Guzi, J.H. Brock, G. Brundu, C.C. Child, C. Daehler, and P. Pyšek), pp. 73–91. Backhuys Publishers, Leiden, Netherlands.

CABI (2018) *Invasive species compendium.* CAB International, Wallingford, UK.

Gan, J., Miller, J.H., Wang, H., & Taylor, J.W. (2009) Invasion of tallow tree into southern US forests: influencing factors and implications for mitigation. *Canadian Journal of Forest Research*, **39**, 1346–1356.

Gucker, C. (2010) Fire Effects Information System (FEIS). Available at: https://www.feis-crs.org/feis/.

Hill, M.J. & Luck, R. (1991) The effect of temperature on germination and seedling growth of temperate perennial pasture legumes. *Australian Journal of Agricultural Research*, **42**, 175–189.

Kalburtji, K.L., Mosjidis, J.A., & Mamolos, A.P. (2007) Effects of day-night temperature combinations under constant day length on emergence and early growth of *Sericea lespedeza* genotypes. *Canadian Journal of Plant Science*, **87**, 77–81.

Loan, A. Van (2006) Japanese climbing fern: The insidious "other" *Lygodium*. *Wildland Weeds*, **9**, 25–27.

Nijjer, S., Lankau, R.A., Rogers, W.E., & Siemann, E. (2002) Effects of temperature and light on Chinese tallow (*Sapium sebiferum*) and Texas sugarberry (*Celtis laevigata*) seed germination. *Texas Journal of Science*, **54**, 63–68.

Orwa, C., Mutua, A., Kindt, R., Jamnadass, R., & Anthony S (2009) Tree Functional and Ecological Databases. Available at: http://www.worldagroforestry.org/output/tree-functional-and-ecological-databases

Qiu, J., Mosjidis, J.A., & Willliams, J.C. (1995) Variability for temperature of germination in *Sericea lespedeza* germplasm. *Crop science*, **35**, 237–241.

You, Y., Liu, H., Wu, R., & Lin, Y. (2008) Effect of low temperature stress on cold resistance of *Cinnamomum camphora* seedling. *Guangdong Agricultural Sciences*, **11**, 23–25.