WILEY Diversity and Distributions

A unified framework to model the potential and realized distributions of invasive species within the invaded range

Tarek Hattab[1] | Carol X. Garzón-López[1] | Michael Ewald[2] | Sandra Skowronek[3] |
Raf Aerts[4] | Hélène Horen[1] | Boris Brasseur[1] | Emilie Gallet-Moron[1] |
Fabien Spicher[1] | Guillaume Decocq[1] | Hannes Feilhauer[3] | Olivier Honnay[4] |
Pieter Kempeneers[5] | Sebastian Schmidtlein[2] | Ben Somers[6] |
Ruben Van De Kerchove[5] | Duccio Rocchini[7,8] | Jonathan Lenoir[1]

[1]UR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, FRE 3498 CNRS-UPJV), Université de Picardie Jules Verne, Amiens, France

[2]Institute of Geography and Geoecology, Karlsruhe Institute of Technology, Karlsruhe, Germany

[3]Institute of Geography, FAU Erlangen-Nuremberg, Erlangen, Germany

[4]Biology Department, KU Leuven, Leuven, Belgium

[5]VITO Flemish Institute for Technological Research, Mol, Belgium

[6]Department of Earth and Environmental Sciences, KU Leuven, Leuven, Belgium

[7]Center Agriculture Food Environment, University of Trento, Michele all'Adige, Italy

[8]Centre for Integrative Biology, University of Trento, Povo, Italy

Correspondence
Tarek Hattab, UR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, FRE 3498 CNRS-UPJV), Université de Picardie Jules Verne, Amiens, France.
Email: hattab.tarek@gmail.com

Abstract

Aim: To propose a species distribution modelling framework and its companion "iSDM" R package for predicting the potential and realized distributions of invasive species within the invaded range.

Location: Northern France.

Methods: The non-equilibrium distribution of invasive species with the environment within the invaded range affects the environmental representativeness of species presence–absence data collected from the field and introduces uncertainty in observed absences as these may either reflect unsuitable sites or be incidental. To address these issues, we here propose an environmental systematic sampling design to collect presence–absence data from the field and a probability index to sort and subsequently separate environmental absences (EAs: reflecting environmentally unsuitable sites) from dispersal-limited absences (DLAs: reflecting sites out of dispersal reach). We first conducted a comprehensive test based on a virtual species to evaluate the performance of our framework. Then, we applied it on different life stages of a non-native tree species (*Prunus serotina* Ehrh.) invasive in Europe.

Results: Regarding the potential distribution, we found higher model performances for both the virtual species (true skill statistics (TSS) > 0.75) and *P. serotina* (TSS ≥ 0.68) after carefully selecting absences with a low probability to be DLAs compared with classical models that incorporate both EAs and DLAs (e.g. TSS = 0.11 for *P. serotina* with 80% of DLAs). On the contrary, both EAs and DLAs as well as dispersal-related covariates were needed to capture the realized distribution of both the virtual species and *P. serotina*.

Main Conclusions: Our framework helps overcoming the conceptual and methodological limitations of the disequilibrium in species' distribution models inherent to invasive species and enables managers to robustly estimate both the realized and potential distributions of invasive species. Although more relevant for modelling the distribution of non-native species, this framework can also be applied to native species.

KEYWORDS
alien species, biological invasions, dispersal limitations, potential niche, realized niche, species distribution modelling, virtual species

1 | INTRODUCTION

Species distribution models (SDMs; see Booth, Nix, Hutchinson, & Jovanic, 1988 for one of the first applications and Guisan & Thuiller, 2005 for a review) provide useful information for managing biological invasions, even after species have established. For instance, SDMs may help identifying priority areas and determining containment boundaries (Robinson, van Klinken, & Metternicht, 2010; Václavík & Meentemeyer, 2009) close to the colonization front where proactive management strategies are needed. The most common SDMs, mainly based on correlational techniques, have provided a popular analytical framework for predicting biological invasions and prioritize locations for early detection and eradication (Booth, Nix, Busby, & Hutchinson, 2014; Gallien, Münkemüller, Albert, Boulangeat, & Thuiller, 2010; Gormley et al., 2011; Jiménez-Valverde et al., 2011). However, when applied specifically to an expanding range of invasive species and especially so at an early invasion stage, SDMs are heavily challenged by the underlying assumption that species distribution should be in a quasi-equilibrium

state with the environment at the invaded range (Gallien, Douzet, Pratte, Zimmermann, & Thuiller, 2012; Guisan & Thuiller, 2005). In other words, it is assumed that the invasive species has already reached all suitable areas and is absent from all unsuitable sites (Figure 1a). However, unoccupied areas in the invaded range may not be due to environmentally unsuitable habitats, but simply due to the fact that the species has not yet had sufficient time to reach these areas (e.g. due to dispersal limitation; Jiménez-Valverde et al., 2011; Uden, Allen, Angeler, Corral, & Fricke, 2015). Consequently, due to a perceived difficulty in interpreting the meaning of real absences from field observations at presumably suitable sites, there is an increased probability of introducing false negative errors into SDMs (Václavík & Meentemeyer, 2009).

In addition to the inherent uncertainties associated with absence data, the level of disequilibrium between the realized and potential distributions also affects the extent to which presence–absence observations provide a representative sample of the environmental domain occupied by the species (Araújo & Pearson, 2005; Václavík & Meentemeyer, 2009; Gallien et al., 2012; Figure 1a). To estimate the
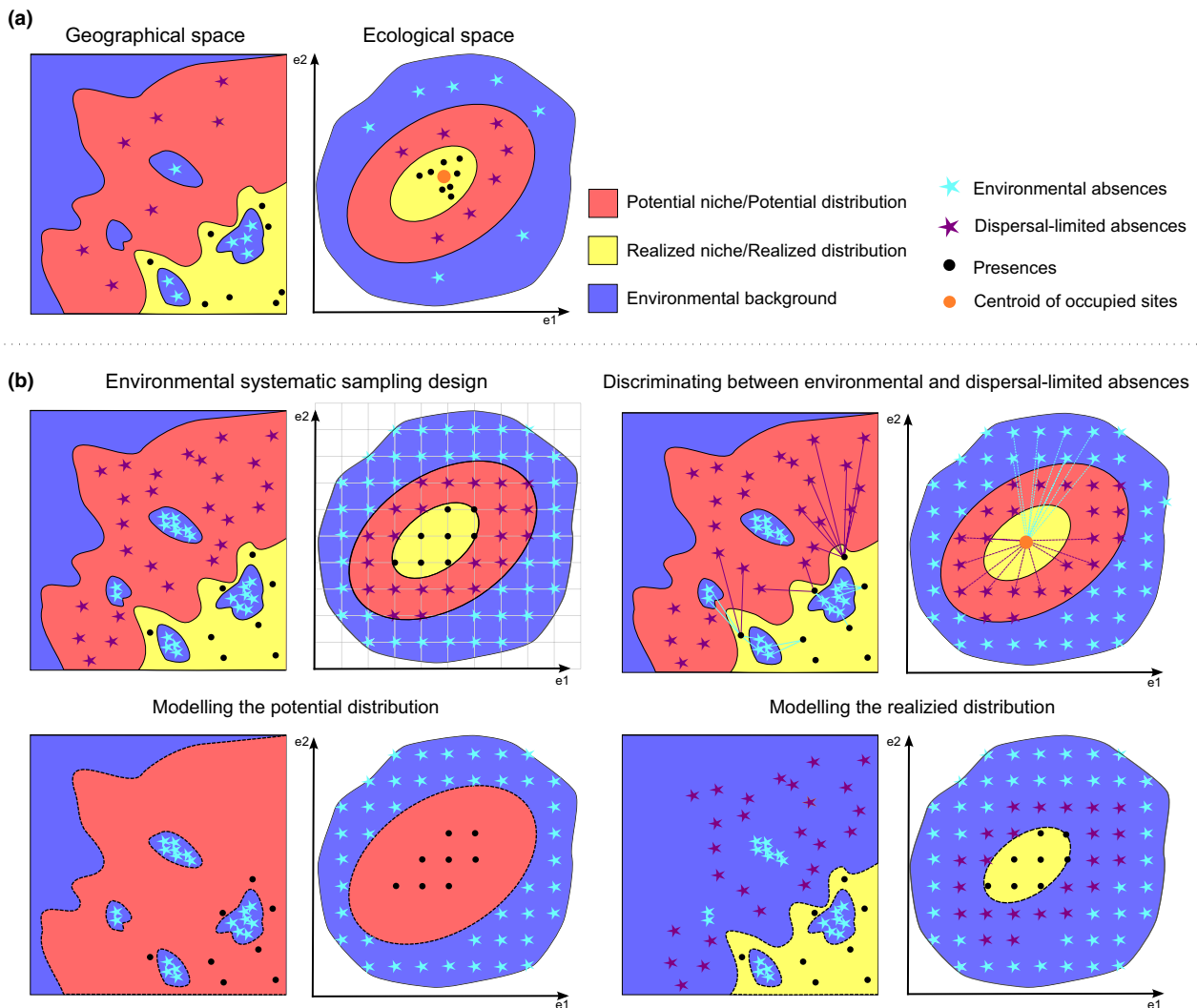

FIGURE 1 Theoretical representation of (a) the disequilibrium in species' distribution within the invaded range and (b) the main steps of our unified framework to model the potential and realized distributions of invasive species within the invaded range based on an environmental systematic sampling design. [Colour figure can be viewed at wileyonlinelibrary.com]

full range of species requirements, some researchers used occurrence data from the native range only (e.g. Gallien et al., 2010; Ibáñez et al., 2009), while others recommended that all data available (from both native and invasive ranges) should be used (e.g. Gallien et al., 2012; Mainali et al., 2015). However, such approaches require good quality data from the native range, which are often not available. Hence, the key challenge for predicting the potential distribution of a non-native species for which only data from the invaded range are available is to handle the uncertainty inherent to absence data while maximizing the level of representativeness of the extent of the ecological niche in the environmental space (Araújo & Pearson, 2005; Robinson et al., 2010; Václavík & Meentemeyer, 2009).

To improve the level of environmental representativeness of species distribution data, former work argued that sampling in a systematic fashion across the environmental space is desired (Fei & Yu, 2016). Thus, collecting distribution data from the invaded range using an environmental systematic sampling design should make it possible to obtain the most complete estimate of a species' potential niche, which is defined as the portion of the fundamental niche space that is potentially available to the species across the studied area (Jackson & Overpeck, 2000). Besides, in regard to absence data uncertainty, previous studies showed that "environmental absences" (EAs; i.e. due to environmentally unsuitable conditions) are required if the goal is to project the species' potential distribution (Lauzeral, Grenouillet, & Brosse, 2012; Lobo, Jiménez-Valverde, & Hortal, 2010; Václavík & Meentemeyer, 2009). On the other hand, dispersal-limited absences (DLAs; i.e. due to dispersal limitation, geographical characteristics or historical events at a priori environmentally suitable sites) when accompanied by predictor variables reflecting dispersal constraints improve the mapping of the species' realized distribution (Lobo et al., 2010; Václavík & Meentemeyer, 2009) which reflects the portion of the potential niche space that is effectively occupied. Thus, it is of utmost importance that modellers provide the right tools to discriminate between EAs and DLAs to accurately map the realized distribution and project the potential distribution (Figure 1b). Both are relevant SDMs' tasks for managing biological invasions (Robinson et al., 2010). For instance, control and eradication efforts should focus on the realized distribution, containment efforts should focus on the interface between the realized and potential distributions, whereas incursion monitoring should focus on the potential distribution (Gormley et al., 2011). The reliability and meaning of SDMs projections for evaluating management options depend on SDMs' ability to make a clear distinction between the potential and realized distributions of an invader, which in turn depends on our own ability to handle the uncertainty inherent to absence data. Hitherto, there is no unified modelling framework to both map the realized distribution and project the potential distribution of invasive species by carefully improving the level of environmental representativeness of species distribution data and cautiously discriminating between EAs and DLAs when calibrating and validating SDMs.

To fill this timely gap, we here propose a modelling framework including: (1) an implementation of an environmental systematic sampling design for collecting distribution data, and (2) a statistical approach to measuring the probability of detecting DLAs within a network of observed presence–absence data and consequently overcome the conceptual and methodological limitations of distinguishing between the potential and realized distributions. Note that our focus here is on real absences from field observations and not on pseudo-absences or background data from random selection procedures (e.g. Barbet-Massin, Jiguet, Albert, & Thuiller, 2012; Mainali et al., 2015; VanDerWal, Shoo, Graham, & Williams, 2009; Wisz & Guisan, 2009), which is a related but different modelling challenge specific to presence-only model. There is a fundamental difference between a dispersal-limited pseudo-absence generated by a random procedure (cf. artificial DLA) and a real DLA collected from a field observation. Therefore, the proposed probability index which relies on presence–absence modelling brings a novelty compared with previous studies that have focused on pseudo-absence selection methods (e.g. Barbet-Massin et al., 2012; Mainali et al., 2015; VanDerWal et al., 2009; Wisz & Guisan, 2009). By using a virtual species for which we know the potential distribution and for which we can simulate different disequilibrium states by manipulating both the colonization rate of the potential range and the level of environmental representativeness of the presence–absence data within the environmental background, we present and test our framework to model both the potential and realized distributions of invasive species. We finally applied this framework on a specific case.

2 | METHODS

2.1 | iSDM: A framework to model the potential and realized distributions

2.1.1 | Collecting distribution data using an environmental systematic sampling design

The first step of our modelling framework involves the use of an environmental systematic sampling design for collecting distribution data (Figure 1b). The aim of this sampling design is to ensure that the survey coverage is representative of all available habitat combinations within the study area as a whole to avoid under- or overestimating the potential distribution within the study area (Albert et al., 2010; Fei & Yu, 2016; MacLeod, 2010). This systematic sampling design consists in reducing the environmental space using an ordination method (the Hill and Smith's method which handle both quantitative and qualitative variables; Hill and Smith (1976)) as a first step. Thereafter, the convex hull of the environmental space is stratified within a regular grid (Figures 1b and 2a) comprising a given number of grid points (the number of grid points corresponds to the desired sample size). The obtained grid represents the perfect configuration required to adequately survey the environmental space in a given study area and for a given sample size. The last step is to seek the closest environmental unit (pixel) to this ideal configuration. This is achieved by searching the nearest neighbour between each grid point and each pixel in the environmental space based on Euclidean distances (Figure 2a). To
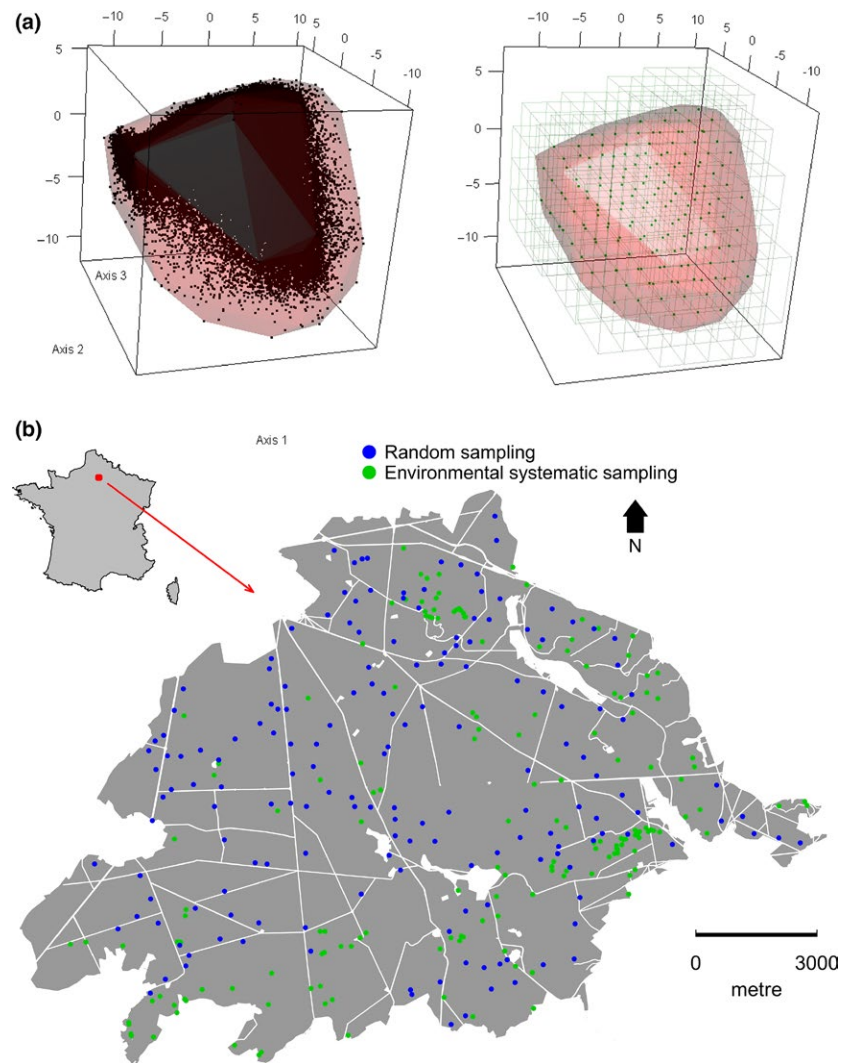
FIGURE 2 Representation of the environmental systematic sampling design within (a) the environmental space and (b) the geographical space across the forest of Compiègne in northern France. Green and blue dots represent plots from the environmental systematic sampling design and field surveys from an earlier field campaign, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

implement this systematic sampling design within the environmental space, we provide the "*eSample*" function in a new R (R Core Team, 2015) package called "iSDM" freely available at https://github.com/TarekHattab/iSDM.

2.1.2 | Discriminating between environmental and dispersal-limited absences (EAs vs. DLAs)

The second step of our modelling framework involves the sorting of absence data to identify the set of absences likely to be limited by dispersal. For these purposes, we developed an index relying on the assumption that a real absence from field observations which is simultaneously located far from the set of observed presences within the geographical space (i.e. outside dispersal reach) but very close to it within the environmental space (i.e. a suitable site outside the realized niche but within the potential niche sensu Jackson and Overpeck (2000)) is likely to be a DLA (Figure 1b). This index goes back to the idea of Acevedo, Jiménez-Valverde, Lobo, and Real (2012) who proposed to reduce the spatial extent of the training area using a geographical criterion to increase the probability of working under a quasi-equilibrium state. However, here we built an index (λ_i)

combining both environmental and geographical criteria to calculate the probability, for any unoccupied site i within the invaded range, that it reflects a DLA (Figure 1b):

$$\lambda_i = \frac{\left(1 - \frac{E_i - \min(E_{ij})}{\max(E_{ij}) - \min(E_{ij})}\right) + \frac{G_i - \min(G_{ij})}{\max(G_{ij}) - \min(G_{ij})}}{2}, \quad (1)$$

where E_i is the Mahalanobis distance between the environmental conditions in the unoccupied site of interest i and the centroid of occupied sites within the environmental space; E_{ij} are all the Mahalanobis distances between the environmental conditions in each of the ij unoccupied sites and the centroids of occupied sites within the environmental space; G_i is the geographical distance between the unoccupied site of interest i and the nearest occupied site; and G_{ij} are all the geographical distances between each of the ij unoccupied sites and their respective nearest occupied sites. Using this additive mathematical formulation, λ_i ranges between 0 and 1, with λ_i values close to 1 indicating a high probability for an unoccupied site to be considered as a DLA. For the computation of distances within the environmental space, we used the Mahalanobis distance because this measure takes into account the correlation among descriptors and is independent of the scales of the descriptors. It measures multivariate environmental

TABLE 1 Predictor variables used in this study. For more information and for maps, see Appendices S1–S3. ONF is the French forests national office

Types of variable	Variable names	Layer types	Resolutions or polygon numbers	Data sources	Data type	Variables used in *Prunus serotina* SDMs
Topographic variables	Elevation (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean + standard deviation
	Slope (degree)	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Eastness	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Northness	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean
	Tangential curvature (1/m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean
Resource availability variables	Insolation time (h)	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Annual global radiation ($Wh.m^{-2}.day^{-1}$)	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean + standard deviation
	Flow accumulation	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Topographic wetness index	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean
Biotic variables	Canopy density	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Minimum canopy height (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean + standard deviation
	Mean canopy height (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Maximum canopy height (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Mean
	5th percentile of canopy height (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	95th percentile of canopy height (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	Standard deviation of canopy height (m)	Raster	0.5 m	LiDAR (ONF)	Numeric	Not used
	NDVI	Raster	1.5 m	SPOT6 satellite image	Numeric	Mean
Soil properties variables	Soil pH	Raster	25 m	Regression kriging	Numeric	Cell values
	Sol type	Vector	7151 polygons	ONF	Factor (6 levels)	Not used
Anthropogenic variables	Stand age	Vector	2846 polygons	ONF	9 intervals	Cell values
	Stand type	Vector	2846 polygons	ONF	Factor (6 levels)	Cell values
	Forest management practices	Vector	2846 polygons	ONF	Binomial	Cell values

similarity between unoccupied and occupied sites and is particularly well suited when the descriptors show multicollinearity (Clark, Dunn, & Smith, 1993; Franklin, 2010). To compute λ_i for each unoccupied site i within a network of presence–absence data, we provide the "*pDLA*" function in the "iSDM" R package.

2.1.3 | Modelling the potential distribution

Once λ_i values have been calculated for all unoccupied sites, we propose to set a λ threshold value beyond which the unoccupied sites are considered as DLAs and thus discarded from both the calibration and validation datasets when modelling the potential distribution (Figure 1b). To set the optimal λ threshold value, we propose to use a comprehensive approach that tests all possible λ threshold values and seeks to select the λ threshold value allowing the optimization of the model's predictive accuracy. This approach can be extended into an iterative ensemble modelling approach somewhat similar to that proposed by Lauzeral et al. (2012). However, unlike Lauzeral et al. (2012) who propose to transform iteratively noisy absences (cf. unoccupied sites predicted to be occupied by an ensemble modelling approach) into presences to mechanistically improve model fit, here we propose to test in parallel all possible λ threshold values corresponding to several careful selections (according to the niche theory) of unoccupied sites that will be discarded (not transformed into presences) from both the calibration and validation datasets. Consequently, in our niche-based approach, each iteration is independent of the previous one, thus eliminating any potential source of error propagation.

2.1.4 | Modelling the realized distribution

To produce models reflecting the realized distribution of the invasion, we propose to use all available absences, both DLAs and EAs, as input

data (Figure 1b) and to also incorporate into the models additional predictor variables describing dispersal-related measures (Meentemeyer, Anacker, Mark, & Rizzo, 2008; Menuz, Kettenring, Hawkins, & Cutler, 2015). Meentemeyer et al. (2008) proposed to quantify the force of invasion (F_i) as a negative exponential dispersal kernel:

$$F_i = \sum_{K=1}^{N} \exp\left(\frac{-d_{ik}}{a}\right), \quad (2)$$

where d_{ik} is the geographical distance between each potential source of invasion k and a target location i. The parameter a modifies the form of the dispersal kernel where low values of a indicate localized dispersal in the vicinity of occupied sites (e.g. autochory), whereas high values indicate potential for dispersal from any occupied site (e.g. allochory). Thus, this dispersal kernel has the advantage of being modified by individual species attributes. The optimal value of a can be selected by trying all possible integer values of a between 0 and 1 and selecting the value that optimizes the model's predictive accuracy (Meentemeyer et al., 2008; Václavík & Meentemeyer, 2009). To create a species-specific dispersal kernel, we provide the "iForce" function in the "iSDM" R package.

## 2.2 | Study site: The forest of Compiègne

To evaluate and apply our modelling framework, we focused on a comprehensive study area located in northern France: the forest

of Compiègne covering 14,417 ha (Figure 2b). This forest was chosen because high-quality presence–absence data on the realized distribution of the American black cherry (*Prunus serotina* Ehrh.; Chabrerie et al., 2007; Chabrerie, Verheyen, Saguez, & Decocq, 2008; Sebert-Cuvillier et al., 2008), a non-native invasive species in Europe (Godefroid, Phartyal, Weyembergh, & Koedam, 2005; Verheyen, Vanhellemont, Stock, & Hermy, 2007), as well as fine-resolution data on a wide range of predictor variables are readily available to implement our modelling framework on a real case study (see Table 1 for the full list of variables and Appendices S1 and S2 for more details).

2.3 | Simulations using a virtual species: A comprehensive test

To evaluate the performance of our modelling framework, we created a virtual species for which we know exactly how the potential distribution should look like and for which we can easily increase the proportion of EAs versus DLAs as well as manipulate the realized distribution to simulate different scenarios of invasion and niche filling. For the sake of simplicity, we used only two predictor variables at 25-m resolution: elevation and soil pH (see Appendix S3 for maps). Focusing on a bivariate environmental space confined to the elevation and soil pH conditions available within the study area, the potential niche (sensu Jackson and Overpeck (2000)) of the virtual species

FIGURE 3 Impacts of an increasing proportion of dispersal-limited absences (DLAs) used to calibrate models of the potential distribution of the virtual species on several statistics measuring model performances: true skill statistic; Pearson's correlation coefficient; specificity; and sensitivity. Results from models based on environmental absences only (EAs: set of absences below the optimal λ threshold value), all absences (EAs + DLAs) and a weighting scheme of each absence ($1/\lambda_i$) are displayed for comparison purposes. Outputs from both the additive formulation (see Equation 1) and a multiplicative formulation (see Appendix S7) of λ_i are displayed separately. The shaded bands represent the range of variability among the eight used modelling algorithms (see Methods section), while the plain and dotted lines represent the mean values. [Colour figure can be viewed at wileyonlinelibrary.com]
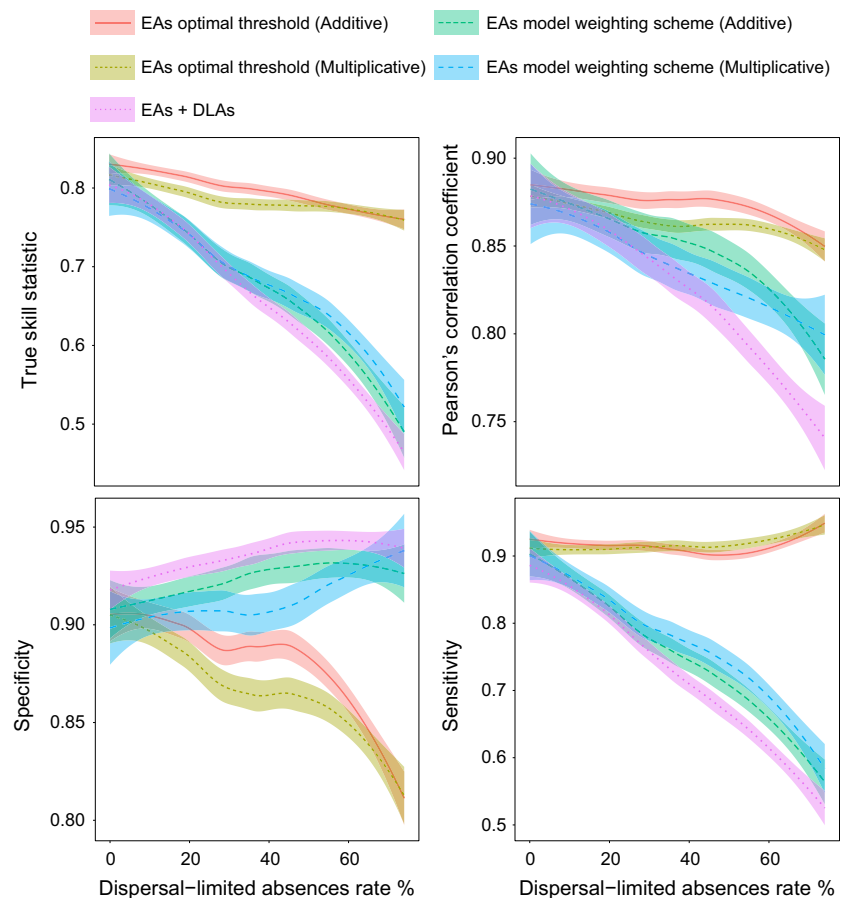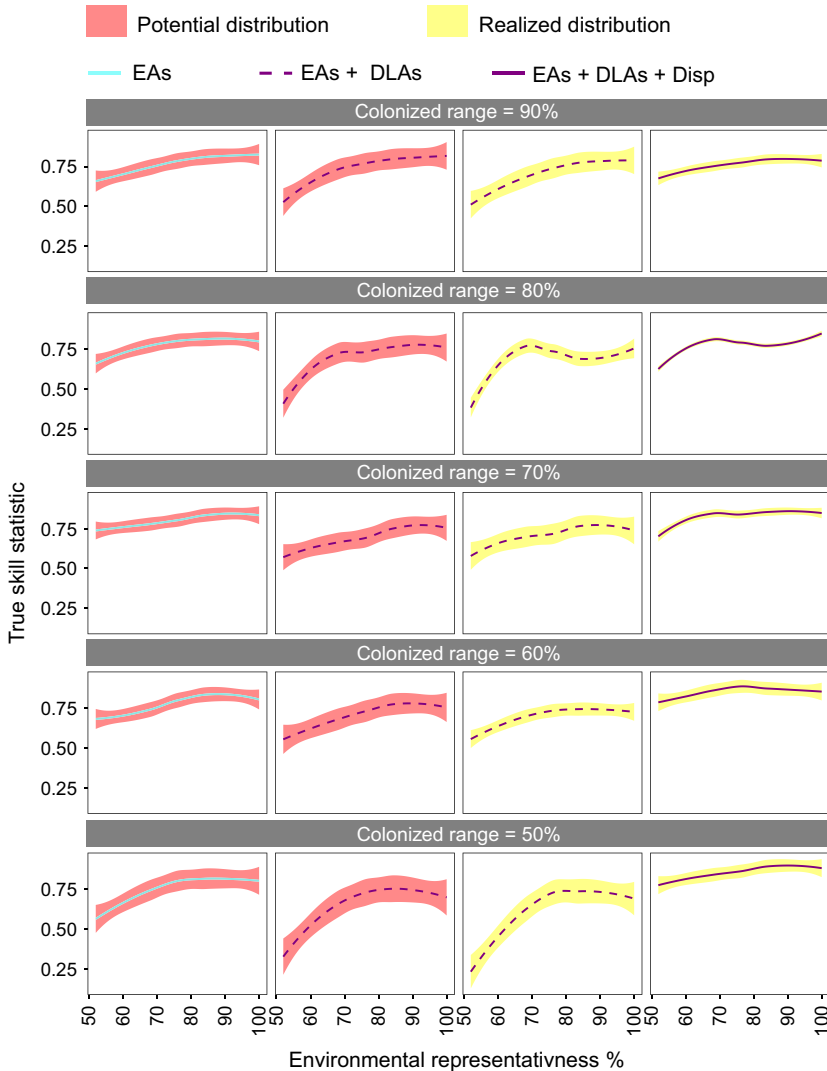
FIGURE 4 Respective impacts of five different scenarios of invasion rate (see Appendix S4) and an increasing level of environmental representativeness of the set of presence–absence data (see Appendix S5) on the true skill statistic (TSS). The shaded coloured bands and the plain/dotted lines represent the range of variability of TSS values and the mean TSS values, respectively, among the eight algorithms (see Methods section) used to model the potential and realized distributions of the virtual invasive species. For comparison purposes, we displayed results from models based on either: environmental absences (EAs: set of absences below the optimal λ threshold value; see Equation 1) only; both environmental and dispersal-limited absences (EAs + DLAs); or all absences as well as dispersal-related covariates (EA + DLAs + Disp). [Colour figure can be viewed at wileyonlinelibrary.com]

was generated using the "virtualspecies" package (Leroy, Meynard, Bellard, & Courchamp, 2015) in R. A simple sigmoid function and a normal function were used to characterize the response curves of the virtual species along the elevation and soil pH gradients, respectively (Appendix S4).

The simulated probabilities of occurrence were transformed into patterns of presence and absence (again not pseudo-absence) using the probabilistic approach described in Meynard and Kaplan (2013) to avoid convergence issues (Meynard & Kaplan, 2013). This simulated presence–absence map represents the potential distribution of the virtual species within the study area. To be able to test our modelling framework, we also had to simulate presence–absence maps representing the realized distribution of the virtual species under different disequilibrium scenarios. As a baseline disequilibrium scenario, we considered that the virtual species invaded, from the south of the study area, only 50% of its potential distribution (see Appendix S5). Using these two (cf. potential and realized) simulated presence–absence maps, we challenged the modelling performances of SDMs against two acknowledged issues in the case of invasive species: (1) an increase in the proportion of DLAs, and (2) a decrease in the level

of environmental representativeness of the presence–absence data within the studied environmental background. For each SDM, we generated a calibration dataset and a validation dataset, each comprising 300 randomly selected presences and 300 randomly selected absences. Although absence data are generated by a random selection procedure in this specific case of the virtual species, it still fundamentally differs from pseudo-absence data as we know the potential niche of the virtual species and thus we know which absences are limited by dispersal. Within each calibration set of 300 absences, both EAs and DLAs were selected. All presences in the calibration dataset were selected so that each presence belongs to both the potential and realized distributions. To assess the impact of an increasing rate of DLAs on the predictive performances of SDMs when modelling the potential distribution, we generated 38 presence–absence calibration datasets by increasing the proportion of DLAs from 0% up to 75% by 2% increments. These SDMs were validated using 38 presence–absence validation datasets reflecting the potential distribution.

To assess the impact of a decreasing rate of disequilibrium between the realized and potential distributions, we first created four hypothetical scenarios of increasing invasion rates, in addition to the baseline
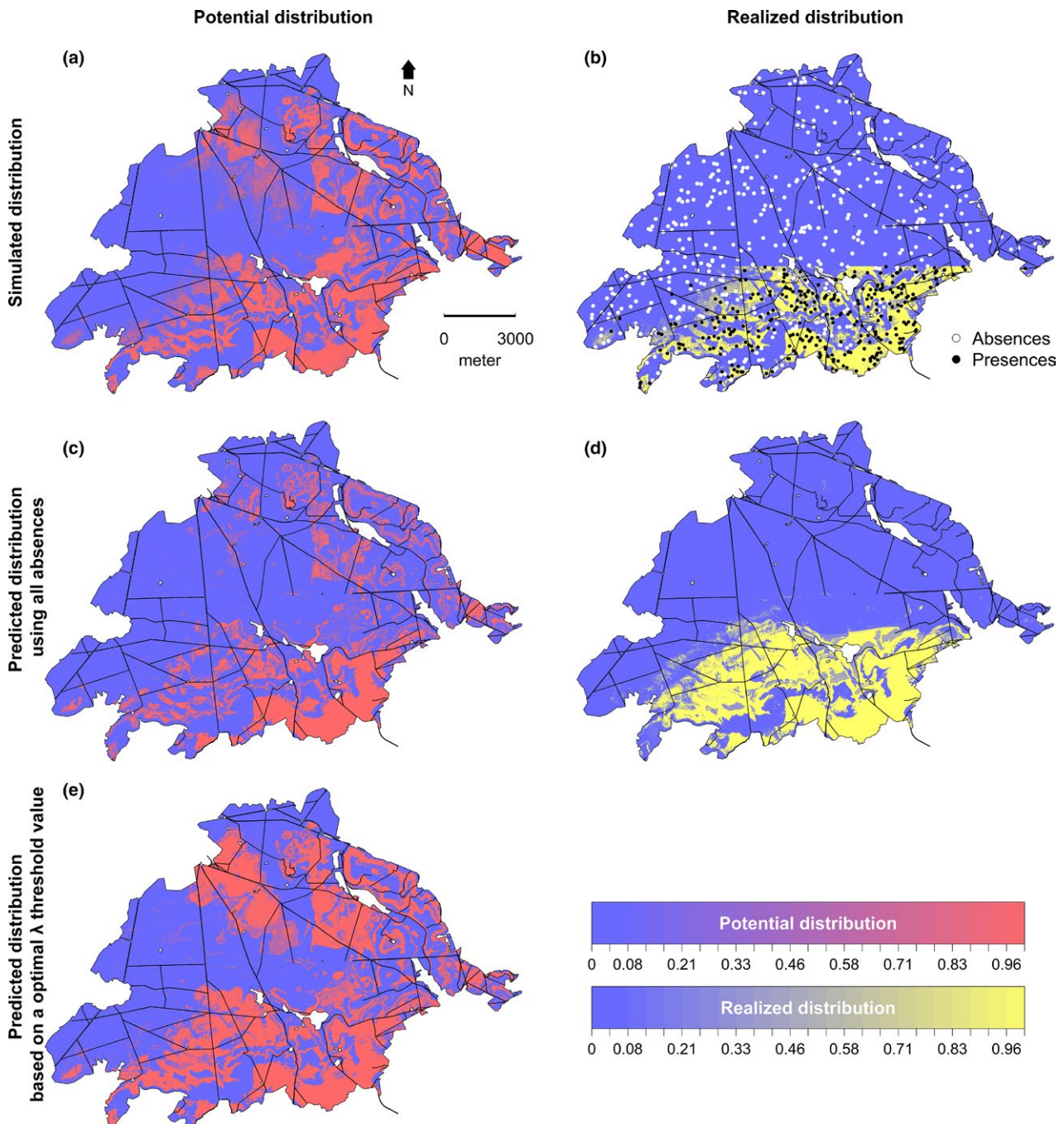
FIGURE 5 Maps of simulated (a, b) and predicted (c, d, e) potential (a, c, e) and realized (b, d) distributions of the virtual species. In this example, we considered an invasion rate of 50%. [Colour figure can be viewed at wileyonlinelibrary.com]

scenario (50% invasion rate), by considering that the virtual species has colonized 60%, 70%, 80% and 90% of its potential range (Appendix S5). Additionally, we varied the disequilibrium state between the realized and potential distributions by increasing the environmental representativeness of the set of presence–absence data within the studied environmental background (cf. the set of elevation and soil pH conditions that is available within the entire study area). For each of the five invasion scenarios, 11 calibration presence–absence datasets were generated by increasing the level of environmental representativeness from 50% up to 100% by 5% increments (see Appendix S6). For each of the 55

disequilibrium scenarios, 110 different validation datasets (55 reflecting the realized distribution and 55 reflecting the potential distribution) were generated, each comprising 300 presences and 300 absences.

For any given set of presence–absence data, and to account for the effect of the choice of the modelling algorithm, we used: generalized linear models; generalized additive models; general boosting methods; classification tree analyses; artificial neural networks; flexible discriminant analyses; multivariate adaptive regression splines; and random forests. To implement these modelling techniques, we used the "biomod2" package in R (Thuiller, Lafourcade, Engler, & Araújo, 2009).

When modelling the potential distribution, we either used: (1) all 300 absences without distinguishing between EAs and DLAs (cf. typical SDM), or (2) only the subset of absences likely to be EAs (cf. below the optimal λ threshold value computed in Equation 1). In addition to the additive formulation described in Equation (1) for calculating λ_i values, we also tested a multiplicative formulation of λ_i (see Appendix S7) and we evaluated the effect of the choice of an optimal λ threshold value instead of a models' weighting scheme (cf. giving more weight to environmental absences using $1/\lambda_i$) on predictive performances. To produce models reflecting the realized distribution, we used either: (1) all absences with incorporating the force of invasion (F_i; see Equation 2) as covariate into our SDMs, or (2) all absences without incorporating F_i (i.e. to assess the relative impact of F_i on the realized distribution). For model evaluation, we used the Pearson's correlation coefficient (between predicted and simulated probabilities of presence), the true skill statistic (TSS) as well as the sensitivity (presences correctly predicted as presences) and the specificity (absences correctly predicted as absences) metrics. To transform occurrence probabilities into presence–absence data, we used the probability threshold that maximized the TSS score of the training samples. Models were evaluated against both the "true" potential and "true" realized distributions as both are known for the virtual species.

2.4 | Application using a real invasive species: The case of *Prunus serotina* Ehrh

To apply our framework on a real case study, we used field observations on the presence–absence of *Prunus serotina* Ehrh. across the forest of Compiègne. Previous studies focusing on *P. serotina* showed that the spatial spread of this long-lived tree species depends on the studied life stage: the invasion range of adult trees is a subset of the invasion range of younger life stages (Sebert-Cuvillier et al., 2007, 2008). Therefore, modelling the potential and realized distribution for the seedling, shrub and tree life stages of *P. serotina* across the forest of Compiègne represents an excellent opportunity to test our framework for different disequilibrium states. Between 2014 and 2015, a total of 166 and 170 plots of 25 m by 25 m were surveyed, following an environmental systematic and a random sampling design (Figure 2), respectively, to collect presence–absence data for *P. serotina*. For the environmental systematic sampling design, we used the three-first ordination axes (Figure 2a) representing all the environmental predictors (except soil pH) described in Appendix S1. The dataset collected randomly was combined with the 166 occurrence data obtained from the systematic sampling design to improve the spatial coverage of the latter and resulting in a final dataset comprising 336 occurrences. At each plot location, the occurrences of *P. serotina* were recorded in three different vegetation layers: herb layer (<0.5 m); shrub layer (0.5–6 m); and tree layer (>6 m).

To model the potential and realized distributions of each of the three life stages, we first aggregated each of the 18 predictors available in a raster format (Table 1) at a 25-m resolution to match the plot size (cf. field surveys). For the set of variables at 0.5-m resolution, we computed the mean and the standard deviation, resulting in total in a

set of 34 quantitative predictor variables. The standard deviation was used as a surrogate to capture the environmental heterogeneity within a given grid cell. The remaining four vector maps were rasterized at the same spatial resolution. Based on this list of 38 predictor variables and previous knowledge on the biology of *P. serotina* within its invaded range (Chabrerie et al., 2007; Godefroid et al., 2005; Verheyen et al., 2007), we selected a set of 15 variables having low pairwise correlations (see Appendix S8 and Table 1) and being physiologically and ecologically meaningful for modelling the distribution of *P. serotina* across the forest of Compiègne.

To calibrate the three SDMs (one for each life stage), we randomly selected 236 plots out of the 336 plots. The remaining 100 plots were used for validation. A random forest analysis was used to calibrate the SDMs, which were subsequently evaluated using the TSS, the area under the curve (AUC) index, the sensitivity metric and the specificity metric. To compute λ_i for each unoccupied site i, we focused on a three-dimensional environmental space derived from the Hill and Smith's ordination method (Hill & Smith, 1976). We used a reduced environmental space given that Mahalanobis distances do not deal with mixed-type variables. Following the iterative approach of our framework, we removed DLAs based on the optimal λ threshold value from the calibration and validation datasets used to model the potential distribution. When modelling the realized distribution, DLAs were kept and three types of dispersal-related variables were incorporated into our SDMs: the force of invasion (by setting iteratively the *a* parameter in Equation 2); the distance from forest tracks; and the distance from the sites of first introduction (we used four alternative introduction points described in Sebert-Cuvillier et al. (2008) to create this map).

3 | RESULTS

3.1 | Evaluation of the framework based on a virtual species

When modelling the potential distribution, we found that models based on EAs solely always performed better than models using all absences (EAs + DLAs), irrespective of: the simulated proportion of DLAs; the considered disequilibrium scenario; or the modelling algorithm used (Figures 3 and 4; Appendix S9). Our simulations also show that setting the optimal λ threshold value based on the additive formulation (cf. Equation 1) offers higher performances than when setting the optimal λ threshold value based on the multiplicative formulation or a weighting scheme ($p \ll .001$; Figure 3). Most important, the TSS and the Pearson correlation values of the models using all absences almost halved, decreasing linearly from 0.82 to 0.48 and from 0.87 to 0.74, respectively, as the rate of DLAs increases from 0% to 75% (Figure 3). In contrast, the TSS and Pearson correlation values of the models using a carefully selected set of absences (EAs only) remained relatively stable (except for the weighting scheme) as the rate of DLAs increases and significantly higher than values obtained from the models based on all absences ($p \ll .001$). This trend was independent to the used modelling algorithm and mostly reflected the

TABLE 2 The true skill statistic (TSS), the area under the curve (AUC), the sensitivity metric and the specificity metric of the modelled distribution of *Prunus serotina* Ehr. within each vegetation layer (tree, shrub, herb) using either: environmental absences (EAs: set of absences below the optimal λ threshold value; see Equation 1) only; environmental and dispersal-limited absences (EAs + DLAs); or all absences as well as dispersal-related covariates (EA + DLAs + Disp)

Layer	Statistics	EAs	EAs + DLAs	EAs + DLAs + Disp
Tree	TSS	0.714	0.112	0.523
	AUC	0.857	0.556	0.760
	Specificity	1.000	0.851	0.957
	Sensitivity	0.714	0.260	0.566
Shrub	TSS	0.789	0.360	0.555
	AUC	0.894	0.680	0.777
	Specificity	1.000	0.842	1.000
	Sensitivity	0.789	0.518	0.555
Herb	TSS	0.680	0.250	0.538
	AUC	0.840	0.625	0.769
	Specificity	1.000	0.750	1.000
	Sensitivity	0.680	0.500	0.538

trend observed for the sensitivity metric (Figure 3). Values for the specificity metric remained relatively high and stable as the rate of DLAs increases (Figure 3).

Irrespective of the set of selected absences (EAs + DLAs vs. EAs), we found higher TSS and Pearson correlation values as the level of environmental representativeness increases from 50% to 100% (Figure 4; Appendix S9). Most important, when modelling the potential distribution, models including all absences had significantly lower TSS and Pearson correlation values than models based on EAs solely, irrespective of the used modelling algorithm and the considered disequilibrium scenario (both environmental representativeness and colonization rate; $p \ll .001$). Overall, the non-removal of DLAs from the calibration and validation datasets implied a decrease in models' sensitivity (i.e. high omission errors; Figure 3) and accordingly an underestimation of the potential distribution (Figure 5).

When modelling the realized distribution, we found that using all absences and accounting for dispersal limitation implied higher performances than models using all absences without accounting for dispersal limitation ($p \ll .001$; Figure 4; Appendix S9). This improvement is all the more important as the level of disequilibrium increases towards low levels of environmental representativeness and invasion rate ($p \ll .001$). Overall, the non-removal of DLAs from the calibration and validation datasets and the integration of dispersal-related variables into the models implied a better estimation of the realized distribution (Figure 5).

3.2 | Application of the framework based on a real case study

Models using all absences without accounting for dispersal limitation showed the worst performances, irrespective of the life stage (cf. disequilibrium state) and the used metrics (Table 2). When applying our framework to the real case of *P. serotina*, we found that 80%, 68% and 46% of absences have a high probability to be DLAs for the tree, shrub and seedling life stages, respectively. By removing these likely DLAs from the calibration and validation datasets, the models' predictive performances increased for all life stages. We found a higher

improvement for the tree life stage, with TSS values increasing from 0.11 to 0.71, than for the shrub and seedling life stages, with TSS values increasing from 0.36 to 0.78 and from 0.25 to 0.68, respectively. This trend was robust to the metric used and mostly reflected the trend observed for the sensitivity metric (Table 2). Models using all absences and accounting for dispersal limitation had higher performances than models using all absences without accounting for dispersal limitation but lower performances than models using environmental absences solely (Table 2). This trend was robust to the metric used and mostly reflected the trend observed for the specificity metric.

Using a standard RGB (red, green, blue) colour space to plot spatial predictions of the potential versus the realized distributions for each life stage separately, we found that the projected potential distribution of the three life stages was substantially similar, while the mapped realized distribution showed nested patterns of distribution with the realized distribution of trees being nested within the one of shrubs and seedlings (Figure 6).

4 | DISCUSSION

4.1 | Modelling the potential distribution

Both the virtual simulation and real case study clearly show that a careful selection of absences having a low probability of being limited by dispersal is critically important for modelling the potential distribution of a species within its invaded range (Figure 3). Models for invasive species have to date generally been trained on the native range (Gallien et al., 2010; Ibáñez et al., 2009), which increases the probability for the focal invasive species to be in a quasi-equilibrium state between its potential and realized distributions. However, when data from the native range are unavailable or costly to obtain, occurrence data from the invaded range have been used (Jiménez-Valverde et al., 2011; Meentemeyer et al., 2008; Václavík & Meentemeyer, 2012). In this situation, the disequilibrium between the potential and realized distributions of the invader makes the use of absence data challenging (Guisan & Thuiller, 2005; Mainali et al., 2015) by producing
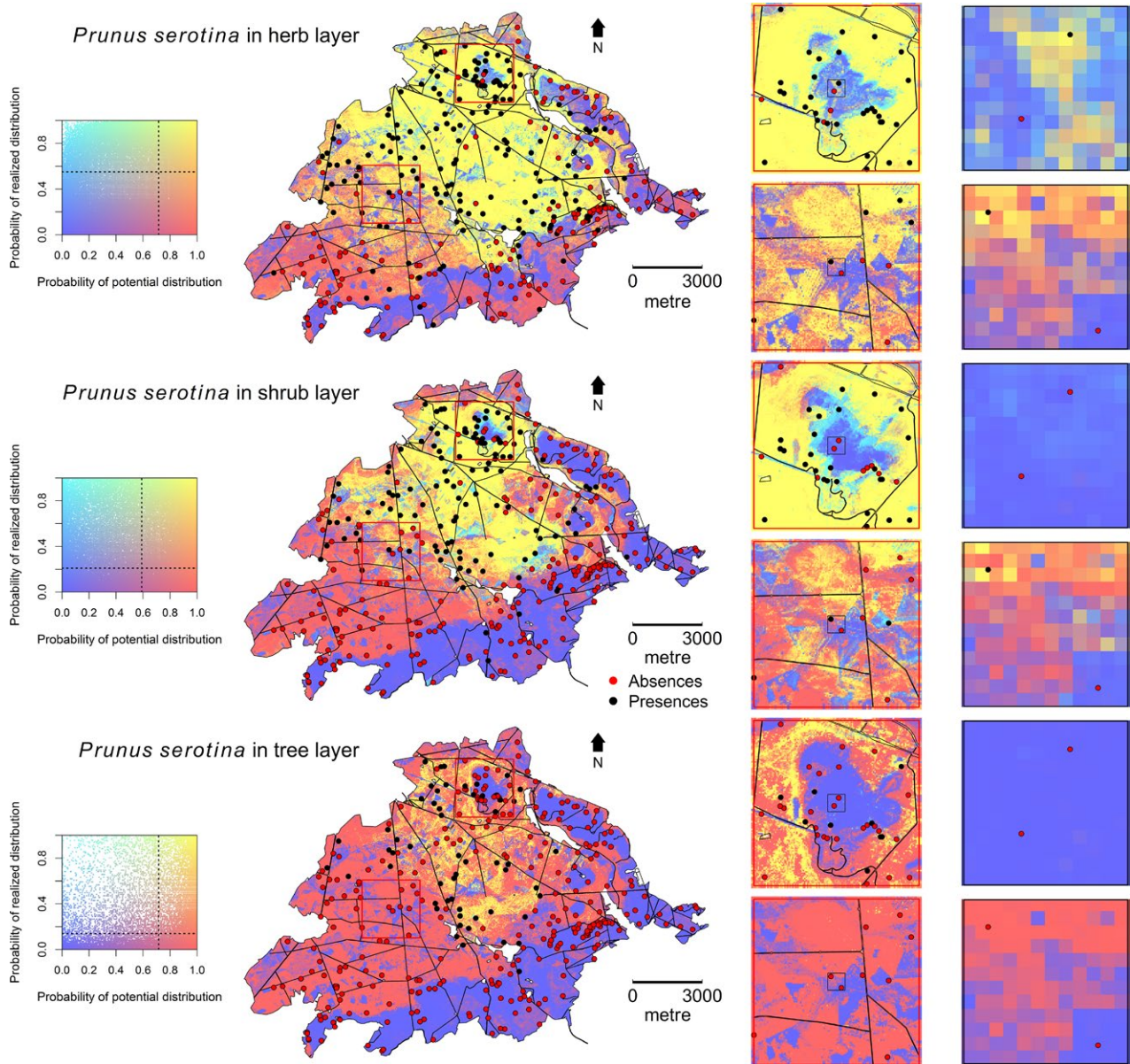
FIGURE 6 Maps of the potential versus the realized distributions for each *Prunus serotina* Ehrh. life stage. Yellowish colours in the three-dimensional RGB (red, green, blue) colour space represent areas with a high likelihood to be already invaded. Reddish colours represent areas not yet invaded but with a high risk of invasion. Dark blue colours represent areas with a low risk of invasion. Light blue colours represent areas likely to be invaded but which should not be given environmental conditions (cf. discrepancies between the realized and the potential distributions suggesting source–sink dynamics sensu Pulliam (2000)). [Colour figure can be viewed at wileyonlinelibrary.com]

a cryptic distinction between the potential and the realized distributions (Václavík & Meentemeyer, 2009). Indeed, our findings suggest that an increase in the proportion of DLAs within the calibration dataset mechanistically decreases the probability to predict species presences successfully (cf. the important decrease in sensitivity observed in Figure 3) and therefore underestimates the potential distribution of the invasive species within its invaded range (Figure 5a, c and e). For this reason, the uncertainties associated with absence data usually lead researchers to simply ignore observed absences and use background absences instead (Gormley et al., 2011; Steiner et al., 2008). Unfortunately, these methods are not able to reliably

estimate the species prevalence (because false zeros of background absences do not have the statistical significance of true absences; (Hastie & Fithian, 2013)) and tend to overestimate the range of invasions (Engler, Guisan, & Rechsteiner, 2004; Václavík & Meentemeyer, 2009). Consistently, our simulations also show that the improvement in projecting the potential distribution after removing likely DLAs is all the more important as the level of invasion rate decreases (cf. the 50% invasion rate in Figure 4). This means that the lower the invasion rate is, the higher the probability to incorporate DLAs is. Coherently, we found a higher proportion of likely DLAs for *P. serotina* in the tree layer (low invasion rate) than in the shrub or in the herb layers (higher

invasion rates). In such situations, removing likely DLAs is of utmost importance to improve projections of the potential distribution. Based on our results, we hence argue that the identification of DLAs is a critical step not only for accurate calibration and validation but also for ecologically meaningful conceptualization of SDMs for biological invasions (Jiménez-Valverde et al., 2011; Lobo et al., 2010; Václavík & Meentemeyer, 2009).

By manually decreasing the environmental representativeness of the presence–absence data within the studied environmental background, we found a decrease in model performances when modelling the potential distribution (Figure 4; Appendix S9). Thus, while our modelling framework allows detecting and excluding likely DLAs to improve model projections for the potential distribution, we also show that it will perform better if the environmental representativeness of the sampled set of presence–absence data is sufficient (i.e. a plateau is reached at approximately 70% of environmental representativeness in Figure 4). It is indeed important to ensure that a sufficient portion of the environmental domain occupied by a given species is sampled when modelling its potential distribution (Mainali et al., 2015). Accordingly, we recommend the use of the "*eSample*" function in the "iSDM" R package before going to the field for setting up a systematic sampling design within the environmental space. Fei and Yu (2016) have shown that field surveys sampled in a systematic fashion across the environmental space provide the best environmental representativeness of occurrence data used in SDMs. Besides, Albert et al. (2010) showed that a systematic sampling design within the environmental space provides an unbiased and high precision estimate of the environmental centroid of occupied sites. Therefore, such a sampling design allows a better estimate of the probability of detecting DLAs (Figure 1). Alternatively, our framework also allows combining presence data from both the native and the invaded ranges (if such data are available) when calculating the probability of detecting DLAs (see the documentation of the "*pDLA*" function in the "iSDM" R package) and when calibrating SDMs (Gallien et al., 2012; Ibáñez et al., 2009; Mainali et al., 2015; Steiner et al., 2008; Uden et al., 2015).

4.2 | Mapping the realized distribution

Importantly, our modelling framework also allows mapping the realized distribution, by using both EAs and DLAs and incorporating dispersal-related covariates in the model. Our results based on simulations from a virtual species clearly show the importance of incorporating dispersal-related covariates to improve model performance (Figure 4) when the aim is to map the realized distribution. This improvement was more pronounced with decreasing level of invasion, even when the environmental representativeness was low (cf. TSS values constantly above 0.8 from 50% to 100% of environmental representativeness at the 50% invasion rate in Figure 4). This suggests that dispersal-related variables play a larger role in mapping the realized distribution of an invasive species within its invaded range than environmental variables particularly when the invasion rate decreases. Coherently, we found a greater increase in the TSS and AUC values when incorporating dispersal-related covariates in models mapping

the realized distribution of *P. serotina* within the tree layer (+0.41 for TSS and +0.20 in AUC) than in models mapping the realized distribution of *P. serotina* within the shrub (+0.19 in TSS and +0.10 in AUC) or herb (+0.29 in TSS and +0.15 in AUC) layers (Table 2). Hence, it is crucial to incorporate dispersal-related variables into SDMs to avoid overestimating the actual invaded range at a given time (Václavík & Meentemeyer, 2009).

4.3 | Potential caveats

Our unified framework to separately model the potential distribution and map the realized distribution does not take into account the case of methodological absences sensu Lobo et al. (2010): the species is present but not detected by the observer. Methodological absences are also important to take into account in SDMs as it may blur the projections of both the potential and realized distributions (Lauzeral et al., 2012). Thus, it is key to apply our unified framework only to easily detectable species, like trees, for which the perception of the realized distribution can be sampled with certainty. In situations of imperfect species detection, the iterative ensemble modelling approach proposed by Lauzeral et al. (2012) can be used to detect and exclude non-environmental absences (cf. both dispersal-limited and methodological absences). Unfortunately, this approach can only be used to model the potential distribution of species difficult to detect but not to map their realized distribution.

5 | CONCLUSIONS

SDMs calibrated without distinguishing between EAs and DLAs and without incorporating dispersal-related covariates overestimate the realized distribution and underestimate the potential distribution of invasive species within their invaded ranges, thereby misdirecting management actions and policy development. Underestimating the potential distribution may result in invasion going unnoticed until the species is well established and thus may put containment and incursion monitoring programs at risk, while overestimating the realized distribution may make costly control and eradication actions unnecessary. The framework we propose enables managers to robustly estimate both the realized and potential distributions of established invasive species within a single unified map (see our "*plotRPD*" function in the "iSDM" R package). Such a map allows to better design management strategies and to detect areas where: (1) the risk of invasion is very low (dark blue colours in Figure 6); (2) invasions have already happened (yellowish colours in Figure 6); (3) the risk of invasion is very high (reddish colours in Figure 6); and (4) there is a discrepancy between the projections of the potential and realized distributions (light blue colours in Figure 6). Interestingly, this discrepancy is more frequent at the juvenile life stages of *P. serotina* and shows that seedlings or saplings may have invaded areas which should not be invaded given environmental conditions only but which are very close to suitable areas. We assume that using dispersal-related covariates in models of the realized distribution but not in models of the potential

distribution may explain this pattern and thus suggests source–sink dynamics (Pulliam, 2000). Noteworthy, our unified framework can also be applied to native species, especially endemics and specialist species, and could yield more efficient results given that the disequilibrium between the realized and the potential distributions is not an issue of specific concern for introduced species but is of far more general concern for all species (Menuz et al., 2015).

AUTHOR CONTRIBUTIONS

T.H. and J.L. conceived the ideas; T.H., M.E., S.S., R.A., H.H., B.B., E.G.M., F.S. and J.L. collected the data; T.H. analysed the data; T.H. led the writing of the manuscript with J.L.; all authors contributed to the final version of the manuscript.

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.

Albert, C. H., Yoccoz, N. G., Edwards, T. C., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). Sampling in ecology and evolution–bridging the gap between theory and practice. *Ecography*, 33, 1028–1037.

Araújo, M. B., & Pearson, R. G. (2005). Equilibrium of species' distributions with climate. *Ecography*, 28, 693–695.

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.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: The first species distribution modelling package, its early applications and relevance to most current MAXENT studies. *Diversity and Distributions*, 20, 1–9.

Booth, T. H., Nix, H. A., Hutchinson, M. F., & Jovanic, T. (1988). Niche analysis and tree species introduction. *Forest Ecology and Management*, 23, 47–59.

Chabrerie, O., Roulier, F., Hoeblich, H., Sebert-Cuvillier, E., Closset-Kopp, D., Leblanc, I., … Decocq, G. (2007). Defining patch mosaic functional types to predict invasion patterns in a forest landscape. *Ecological Applications*, 17, 464–481.

Chabrerie, O., Verheyen, K., Saguez, R., & Decocq, G. (2008). Disentangling relationships between habitat conditions, disturbance history, plant diversity, and American black cherry (*Prunus serotina* Ehrh.) invasion in a European temperate forest. *Diversity and Distributions*, 14, 204–212.

Clark, J. D., Dunn, J. E., & Smith, K. G. (1993). A multivariate model of female black bear habitat use for a geographic information system. *The Journal of Wildlife Management*, 519–526.

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41, 263–274.

Fei, S., & Yu, F. (2016). Quality of presence data determines species distribution model performance: A novel index to evaluate data quality. *Landscape Ecology*, 31, 31–42.

Franklin, J. (2010). *Mapping species distributions: Spatial inference and prediction*. Cambridge: Cambridge University Press.

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N. E., & Thuiller, W. (2012). Invasive species distribution models–how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography*, 21, 1126–1136.

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.

Godefroid, S., Phartyal, S. S., Weyembergh, G., & Koedam, N. (2005). Ecological factors controlling the abundance of non-native invasive black cherry (Prunus serotina) in deciduous forest understory in Belgium. *Forest Ecology and Management*, 210, 91–105.

Gormley, A. M., Forsyth, D. M., Griffioen, P., Lindeman, M., Ramsey, D. S., 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.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8, 993–1009.

Hastie, T., & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 36, 864–867.

Hill, M., & Smith, A. (1976). Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25, 249–255.

Ibáñez, I., Silander, J. A. Jr, Wilson, A. M., LaFleur, N., Tanaka, N., & Tsuyama, I. (2009). Multivariate forecasts of potential distributions of invasive plant species. *Ecological Applications*, 19, 359–375.

Jackson, S. T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, 26, 194–220.

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

Lauzeral, C., Grenouillet, G., & Brosse, S. (2012). Dealing with noisy absences to optimize species distribution models: An iterative ensemble modelling approach. *PLoS ONE*, 7, e49508.

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2015). virtualspecies, an R package to generate virtual species distributions. *Ecography*, 38, 001–009.

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

MacLeod, C. D. (2010). Habitat representativeness score (HRS): A novel concept for objectively assessing the suitability of survey coverage for modelling the distribution of marine species. *Journal of the Marine Biological Association of the United Kingdom*, 90, 1269–1277.

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

Meentemeyer, R. K., Anacker, B. L., Mark, W., & Rizzo, D. M. (2008). Early detection of emerging forest disease using dispersal estimation and ecological niche modeling. *Ecological Applications*, 18, 377–390.

Menuz, D. R., Kettenring, K. M., Hawkins, C. P., & Cutler, D. R. (2015). Non-equilibrium in plant distribution models–only an issue for introduced or dispersal limited species? *Ecography*, 38, 231–240.

Meynard, C. N., & Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. *Journal of Biogeography*, 40, 1–8.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters*, 3, 349–361.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robinson, T. P., van Klinken, R. D., & Metternicht, G. (2010). Comparison of alternative strategies for invasive species distribution modeling. *Ecological Modelling*, 221, 2261–2269.

Sebert-Cuvillier, E., Paccaut, F., Chabrerie, O., Endels, P., Goubet, O., & Decocq, G. (2007). Local population dynamics of an invasive tree species with a complex life-history cycle: A stochastic matrix model. *Ecological Modelling*, 201, 127–143.

Sebert-Cuvillier, E., Simon-Goyheneche, V., Paccaut, F., Chabrerie, O., Goubet, O., & Decocq, G. (2008). Spatial spread of an alien tree species in a heterogeneous forest landscape: A spatially realistic simulation model. *Landscape Ecology*, 23, 787–801.

Steiner, F. M., Schlick-Steiner, B. C., VanDerWal, J., Reuther, K. D., Christian, E., Stauffer, C., … Crozier, R. H. (2008). Combined modelling of distribution and niche in invasion biology: A case study of two invasive Tetramorium ant species. *Diversity and Distributions*, 14, 538–545.

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

Uden, D. R., Allen, C. R., Angeler, D. G., Corral, L., & Fricke, K. A. (2015). Adaptive invasive species distribution models: A framework for modeling incipient invasions. *Biological Invasions*, 17, 2831–2850.

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.

Václavík, T., & Meentemeyer, R. K. (2012). Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions*, 18, 73–83.

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.

Verheyen, K., Vanhellemont, M., Stock, T., & Hermy, M. (2007). Predicting patterns of invasion by black cherry (*Prunus serotina* Ehrh.) in Flanders (Belgium) and its impact on the forest understorey community. *Diversity and Distributions*, 13, 487–497.

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

BIOSKETCH

Tarek Hattab received his PhD in Ecology at the University of Montpellier 2, France; he currently holds a postdoctoral researcher position at the University of Picardie Jules Verne, France. His research focuses on the application of macroecological and spatial ecology concepts to explore patterns and processes of biodiversity from local to global scales.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.