The importance of biotic interactions in distribution models depends on the type of ecological relations, spatial scale and range.

Merijn Moens\textsuperscript{1}, Jacobus Biesmeijer\textsuperscript{1}, Elaine Huang\textsuperscript{2}, Nicolas Vereecken\textsuperscript{3}, and Leon Marshall\textsuperscript{4}

\textsuperscript{1}Naturalis Biodiversity Center
\textsuperscript{2}Wageningen University & Research
\textsuperscript{3}ULB Université Libre de Bruxelles
\textsuperscript{4}Universite de Namur

November 7, 2022

Abstract

Classical Species Distribution Models are primarily based on climate, land use and other abiotic variables. Despite recent studies showing that biotic information can play an important role in shaping the distribution of species even at large scales, results are not always consistent among studies and the underlying factors that influence the importance of this biotic information to the models, are unclear. To address this address gap, we evaluated how different factors affect the importance of biotic interactions in shaping species distributions, using fine-scale data from plant-pollinator and parasitic interactions in the Netherlands. We found that the models of wild bees improved, when their biotic interaction was included, and the model performance improved the most for parasitic bees. Taxonomic level, resolution and distribution range of the interacting species and degree of specialization of the modelled species all affected the importance of the biotic interactions to the models.

The importance of biotic interactions in distribution models depends on the type of ecological relations, spatial scale and range.

Short running title: Factors influencing the importance of biotic interactions in distribution models.

Merijn Moens\textsuperscript{1,2}, Jacobus C. Biesmeijer\textsuperscript{1,2}, Elaine Huang\textsuperscript{3}, Nicolas J. Vereecken\textsuperscript{4}, Leon Marshall\textsuperscript{1,4}

Merijn Moens (https://orcid.org/0000-0002-7503-6572) (merijn.moens@naturalis.nl), Jacobus C. Biesmeijer (https://orcid.org/0000-0003-0328-0573) (koos.biesmeijer@naturalis.nl), Elaine Huang (elaine.huang@wur.nl), Nicolas J. Vereecken (https://orcid.org/0000-0002-8858-4623) (nicolas.vereecken@ulb.be), Leon Marshall (https://orcid.org/0000-0002-7819-7005 ) (Leon.Marshall@ulb.be)

1. Naturalis Biodiversity Center, PO Box 9517, 2300 RA Leiden, The Netherlands 2. Institute of Environmental Sciences (CML), Leiden University, Einsteinweg 2, 2333 CC Leiden, The Netherlands 3. – 4. Agroecology Lab, Interfaculty School of Bioengineering, Université libre de Bruxelles (ULB), Boulevard du Triomphe CP 264/2, B-1050 Brussels, Belgium

Authorship: MM, EH, KB and LM designed and performed the study. MM performed the Species Distribution Models. MM and EH performed the Generalized Linear Models. MM, EH, KB and LM analyzed the output of the models. MM wrote the first version of the manuscript and all authors contributed substantially to revisions. Data accessibility: The data that supports the results will be shared in a Dryad public repository. The DOI will be included in the article. Targeted article type: letter Number of words: 294 words in abstract and 4705 words in main text – 6 figures in main text. – 86 references. Keywords:
species distributions, ecological niche models, biotic interactions, spatial resolution, plant-pollinator interac-
tions, specialists, generalists, parasites, null models, evaluation measures. Telephone number corresponding
author: +31 6 15887829

Abstract

Classical Species Distribution Models are primarily based on climate, land use and other abiotic
variables. Despite recent studies showing that biotic information can play an important role in
shaping the distribution of species even at large scales, results are not always consistent among
studies and the underlying factors that influence the importance of this biotic information to
the models, are unclear. To address this knowledge gap, we evaluated how different factors
affect the importance of biotic interactions in shaping species distributions, using fine-scale
data from plant-pollinator and parasitic interactions in the Netherlands. We found that the
models of wild bees improved, when their biotic interaction was included, and the model
performance improved the most for parasitic bees. Taxonomic level, resolution and distribution
range of the interacting species and degree of specialization of the modelled species all affected
the importance of the biotic interactions to the models.

Introduction:

Over the last thirty years, there has been a massive increase in the number of published studies using
species distribution models (SDMs) (reviewed in Lobo et al. 2010 and Melo-Merino et al. 2020). Species
distribution models are used to identify areas of potentially suitable habitat by linking species occurrences
to environmental variables (Loyola et al., 2012; Silva et al., 2014). These predictions of suitable habitats
have many applications (Elith & Leathwick 2009), including: the estimation of potential distribution under
different climate change scenarios (Marshall at al. 2018; Lima et al. 2020), the estimation of suitable areas for
a species (Suzuki-Ohno et al. 2017) and assessing the potential invasiveness of an exotic species (Srivastava
et al. 2019).

A key to understanding distribution patterns in insects lies in understanding their relationships with other
organisms, e.g. pollinators and their floral resources (Willmer 2011). This biotic information is rarely inclu-
ded directly in distribution models, despite the fact that biotic interactions can theoretically improve our
understanding and predictions of the distribution of species through different mechanisms (Wisz et al. 2013).
Previous studies showed an improvement in the statistical performance of spatial models when including
parasitic (Mathieu-Begne et al. 2021), facilitative (Heikkinnen et al 2007), resource-consumer (Kissling et
al. 2007; Bateman et al. 2012; Roslin et al. 2017; Atauchi et al. 2018; Herrera et al. 2018), competitive
(Leach et al 2016; Mpakairi et al. 2017) and plant-pollinator interactions (Araujo & Luoto 2007; Espindola
& Pliscoff 2019; Kass et al. 2020). However, the improvements made by including biotic interactions depend
on the spatial scale of the biotic variable in the model (Heikkinen et al 2007), and correlative relationships
observed in models do not always represent biotic interactions (Giannini et al. 2013).

The biotic variable can be included in the SDM as an explanatory variable (Araujo & Luoto 2007; Kass
et al. 2020) and it can also be introduced as either the raw distribution (Giannini et al. 2013; Leach et
al. 2016) or a modelled distribution of the species that the modelled species interacts with (Bateman et
al 2012; Giannini et al. 2013) (hereafter “interacting species”). The biotic variable can also be introduced as
an approximation of the intensity of the interaction, e.g. the genomic background of parasite hosts can
help identify populations with resistance genes (Mathieu-Begne et al. 2021), distance towards sighting of a
competitor for competition (Mpakairi et al. 2017) or the distribution of diet resources for resource-consumer
interactions (Araujo et al. 2014).

Besides the methodological considerations, spatial resolution may strongly affect the contribution of biotic
interactions to modelled distribution patterns (Pearson and Dawson, 2003; Soberon & Peterson, 2005; Wisz
et al. 2013). Heikkinen et al. (2007) showed that the impacts of facilitation between owls and woodpeckers
are more visible at a resolution of 10 km than 40 km. This is consistent with Pearson and Dawson (2003), who
hypothesized that at broader scales and coarse resolutions, climate variables are more dominant and biotic
interactions less apparent (Heikkinen et al. 2007). However, an obligate parasite with a strong interaction with its host may always be more dependent on its interacting species’ distribution at any resolution. There is insufficient evidence as to how the explanatory power of biotic factors changes with spatial resolution, which is crucial for improving SDMs of species with strong hypothesized biotic interactions.

Whether the interaction is essential for survival depends on factors that include body size, dietary breadth, the distribution of the interacting species and the dependency on one another, e.g. whether the flowers are similarly dependent on the pollinator as the pollinator on the plant. The bee body size shows a strong relation with the foraging distance of different bees (Greenleaf et al. 2007; Kendall et al. 2019 and references therein) and smaller bees with a smaller foraging distance would require their host plant closer to their nest. It has been hypothesized that the dietary breadth of the species could influence the importance of the biotic factor in the models (e.g. specialist vs generalist; Araujo et al. 2014). In the case of bees, it has been shown that the population trend of specialist bees is correlated to the plant that they are dependent on for their pollen (Schepers et al. 2014). We expect that a smaller distribution range of the interacting species would have a higher contribution to the models, as it more likely to be the limiting factor of the modelled species. Specialist bee species have a tendency to decline more than generalist bee species and their decline is correlated to their host plant (Biesmeijer et al. 2006) and this leads us to expect that the specialist species show a higher contribution of the interacting species to their models.

In this paper, we aim to use a priori knowledge to investigate the importance of biotic interactions in species distribution models of bees at different spatial resolutions. Wild bees are a group of well-studied organisms that include species with a great importance to ecosystem resilience and that play a key role in pollination services to wild plants and crops (Kleijn et al. 2015; Senapathi et al. 2015; Weekers et al. 2022). Bees depend on pollen and nectar provided by plants and diets range from narrow (oligolectic bees, using few plant species) to broad (polylectic bees, using many plant species) (Rasmussen et al. 2020). Other species, up to 30%, are cleptoparasitic, meaning they are brood parasites which lay eggs in nests of other bee species (Cardinal et al. 2010). They may have one or multiple host bee species. The Netherlands is a good case study for the effects of biotic interactions on the distribution of wild bees, as there are more than 300 species of wild bees (Reemer 2018) and there is extensive data on plant-pollinator interactions, hosts of cleptoparasitic bees and occurrence data. By integrating knowledge of plant visitation and cleptoparasitic interactions, we aim to (1) assess the importance of biotic factors in explaining distributions of polylectic, oligolectic and cleptoparasitic bees and (2) assess the relative importance of different factors, including flower and host specialization, spatial resolution, taxonomic level, distribution of the interacting species and bee body size (which relates to foraging range) in explaining the contribution of the interacting species to the models. By modelling a large number of bee species and using different input variables and methods, we identified important factors that are related to the implementation of the biotic interactions to the models.

Materials and Methods:

Variables and data

We used climate data from the Koninklijk Nederlands Meteorologisch Instituut (KNMI 2016) from the period of 2000 to 2015 in bioclim format (Fick & Hijmans 2017). A Principal Component Analysis (PCA) was used to transform the 19 bioclim variables in five orthogonal PCA axes that explained more than 90% of the variation. The resolution of the climate variable was 100 m (100 m by 100 m) and all other variables were transformed to the same resolution. The land use data consisted of 15 land use categories from different sources (see table S1 in supporting information; Inter Provinciaal Overleg, 2016; Ministerie van Economische Zaken (EZK), 2015; Centraal Bureau voor de Statistiek (CBS), 2012). Together with the sum of the different land uses per grid cell as an indicator for landscape heterogeneity, this resulted in 16 land use variables. The 8 main aggregated soil types in the Netherlands (table S2; grondsoortenkaart 2006) and the land use data were transformed to a 100 m resolution raster format with percentage cover values and they were not strongly correlated (Spearman’s $p < 0.7$; Dormann et al. 2013).

The bee occurrence data used in this study consist of opportunistic observations of bees and flower vis-
itation records from 2004 to 2019, obtained from the European Invertebrate Survey Netherlands (EIS; http://www.eis-nederland.nl/). Of the more than 300 bee species in the Netherlands we selected 194 species, discarding the species without flower visitation information and with less than fifteen observations. Using the flower visitation data, the most visited plant species and genus were determined. Bees were classified as either oligolectic (includes both monolectic species which visit a single plant taxon and oligolectic species which show a clear preference for a single plant family), polylectic (collect pollen and nectar from various plant taxa) (Rasmussen et al. 2020) and cleptoparasitic bees; these traits were based on a database created for the Status and Trends of European Pollinators (STEP) project and maintained by Stuart Roberts (Potts et al. 2015). For all the 45 modelled oligolectic bees the most visited plant was the plant that they are dependent on for their pollen in the bee trait database (table S3). Since no quantitative interaction data was available for cleptoparasitic bees and their host, we used the occurrences of all the host bee species and genera in the Netherlands listed in the literature (table S4; Peeters & Nieuwenhuijsen 2012). The plant occurrences for the period 2004-2019 were obtained from Dutch National Database of Flora and Fauna (NDFF 2021). The occurrences of the interacting species were kept at a resolution of 100 m or aggregated to a coarser resolution (500 m, 1 km, 5 km and 10 km) and they presented presence/absence, i.e. binary data, similar to Godsoe et al. (2009). Some of the cleptoparasitic bees had multiple hosts and the biotic interaction variable map had a presence if any of the host species or genera was present.

Species distribution modelling

For the modelling of the species distributions, we performed all models in R version 4.0.3 (R Core Team 2020) with MaxEnt (version 3.4.1) (Phillips et al. 2006) and dismo (version 1.3-3) (Hijmans et al. 2020) and each model was replicated four times, using the cross-block validation method in the ENMeval package version 2.0.3 (Muscarella et al. 2014). Model evaluation measures were averaged across these 4 models. In total, we developed 68,288 models (see Appendix S1 in supporting information) and an overview can be found in figure 1 and table S5 following ODMap (Zurell et al. 2020). We ran all models in parallel using the package snow version 0.4-3 (Tierney et al. 2018) and the package parallel in base R version 4.0.3 (R Core Team 2020). The performance was compared with the Area Under the Curve (AUC) of the Receiver Operating characteristics curve, which is a threshold independent evaluation method of the performance of the calibration and evaluation dataset (Phillips et al. 2006; Elith et al. 2006). We also looked at the percentage contribution of the variable to the model gain, which is a common measure of variable importance (Bradie & Leung 2017). The degree of overfitting was used to select the appropriate regularization multiplier value of 5 (Appendix S2). The evaluation AUC of the different resolutions were ranked per species to reduce variability between species and find the optimal model settings for the different groups of bees. We made prediction maps of the models with the biotic variable at 1 km for the calculation of the corrected Akaike information criterion (AICc; Burnham & Anderson 2002) and Continuous Boyce Index (CBI; Hirzel et al. 2006). Presence and absence maps were made based on the maximum training sensitivity and specificity threshold (De Barros et al. 2012). The AICc and Continuous Boyce Index are evaluation measures that include a penalty based on the number of variables that is used in the models, unlike the AUC. The CBI, AICc and evaluation AUC were compared between models including biotic factors and excluding biotic factors with a one-sample Wilcoxon signed rank (Wilcoxon 1945). We also compared the percentage contribution of the different variable classes (land use, climate, soil, and biotic variables) and individual variables, using a Kruskal-Wallis H test (Kruskal & Wallis 1952) with a post-hoc Nemenyi test (Sachs 1997).

Null models

In the current study, null models consist of SDMs comparing the addition of the known interacting species of the modelled bee to the addition of any other visited plant or a bee that is parasitized on (fig. 1; Appendix S3). These null models allow us to distinguish the specificity of the relations as opposed to emerging relations due to shared habitats or similar sampling locations (e.g., plants and bees may be sampled in similar areas). The null models with the interacting species at species and genus level resulted in a total of 263,952 models for the flower visiting bees and 21,200 for the cleptoparasitic bees (Appendix S3). The evaluation AUC of the model with the interacting species was ranked among the null models per species and the percentage
rank was compared among groups. We analysed per species the ranking of the model with the interacting species compared to the models with random interactions and calculated the percentage of modelled species that were among the 5% and 25% best performing models.

**Generalized linear models**

To assess the relative importance of biotic variables (flower and host specialization and distribution of the interacting species) and bee body size (which relates to foraging range) for explaining wild bee occurrence in the Netherlands, we developed a series of Generalized Linear Models (GLMs) (fig. 1). For the flower visiting bees, we used the plant observation data, and we calculated the amount of grid cells occupied by each visited plant species. Secondly, we computed a measure of flower specialization, calculated as the diversity of genera visited in the interaction database for every flower visiting bee, using the Shannon-Wiener index (Shannon 1948). Thirdly, we used the information on body size from the bee trait database. The body size was measured using the intertropical distance (ITD, in mm) as a proxy, which is the distance between the wing insertion points (Greenleaf et al., 2007). The cleptoparasitic bees were modelled in a similar way, except that the distribution of the interacting species was calculated from observation data from potential host bees (table S2 in supporting information) and the host specialization was the number of potential hosts in the literature (Peeters & Nieuwenhuijsen 2012). In both cases, the explanatory variables were standardized, centred, and a gamma distribution with an inverse link function was used. The gamma distribution is applicable for situations in which we want to speculate about the response variable without certainty about its distribution (Faraway 2016) and for ecological data with non-zero values (Foster & Bravington 2013). Model performance was quantified with the AICc and the $R^2$, which represents the proportion of the variance in the dependent variable that the model explains. Modelled species that did not find a contribution (e.g. no features of the respective variable present in the model) of their interacting species to their distribution were left out of the analysis. The three explanatory variables resulted in eight possible combinations of variables and we evaluated the models using the AICc as described in Hurvich & Tsai 1989. The GLMs were developed in the stats package in base R version 4.0.3 (R Core Team 2020).

**Results:**

**Influence of spatial resolution and taxonomic level on model performance**

The contribution of the interacting species to the models was ranked highest at 500 m resolution (cleptoparasitic bees) and 1 km (oligolectic and polylectic bees) and at both taxonomic levels (fig. 2A; Appendix S4). The optimal resolution of 500 m for the cleptoparasitic and 1 km for the polylectic bees was significantly higher ($p < 0.05$) compared to the other resolutions except for the difference between 500 m and 1 km, which was not significant at both taxonomic levels (Appendix S4). The oligolectic bees only showed a statistically significant difference between the optimal resolution of 1 km and the extremes (100 m and 10 km).

In general, the contribution of the interacting species to the models was higher when added at the species level for coarser resolutions and at the genus level at finer resolutions. The contribution of the interacting species was higher at the species level at a coarser resolution from 5 km to 10 km and higher at the genus level from 100 m to 1 km (fig. 2B; Appendix S4). This general trend was also shown at a resolution from 100 m to 1 km for the cleptoparasitic bees, oligolectic bees and the polylectic bees and at 5 km to 10 km resolution for the models of the cleptoparasitic bees.

The percentage contribution of the interacting species was higher for the cleptoparasitic bees compared to the oligolectic and polylectic bees. This difference was statistically different for all resolutions of the interacting species (Appendix S4). The oligolectic and polylectic bees only showed a statistically significant difference at a 10 km resolution with a higher contribution of the interacting species for the oligolectic bees.

**The effect of biotic factors on overall model performance**

The models improved significantly when the interacting species was included, for all evaluation measurements: Area Under the Curve (AUC), continuous Boyce index (CBI) and corrected Akaike information criteria (AICc). The difference in these evaluation metrics is the change from models with biotic interactions to
models without biotic interactions. The interacting species was included at a resolution of 1 km, which was on average the optimal resolution for most models (fig. 2A). The models of the cleptoparasitic, oligolectic and polylectic bees all showed a statistically significant increase in evaluation AUC (fig. 3A), evaluation CBI (fig. 3B) and a decrease in calibration and evaluation AICc (fig. 3C; Appendix S5). There were not enough samples in the evaluation data to calculate the AICc for a proportion of the bees (34.02% of the species; 66 species) and only the cleptoparasitic bees showed a significant improvement on the species and genus level (fig. 3D). The biotic interaction had a high contribution to the models of all the functional bee groups relative to the climate, land use and soil variables (fig. 4; Appendix S6). The biotic interaction had the highest contribution to the models compared to the climate, land use and soil variables for the cleptoparasitic bees.

The models of the cleptoparasitic bees showed a statistically significant higher improvement in evaluation AUC, evaluation CBI and in calibration and evaluation AICc compared to the oligolectic and polylectic bees with the interacting species at both taxonomic levels (Appendix S5). The difference in evaluation metrics between oligolectic and polylectic bees was in no case significant.

We found that the relative performance of the interacting species compared to any randomly chosen interaction was the highest for the cleptoparasitic, oligolectic and polylectic bees respectively and the polylectic bees showed a more general preference of biotic interactions than the cleptoparasitic and oligolectic bees. The models with the interacting species (added at species and genus taxonomic level) had an evaluation AUC that scored within the highest 5% best performing models of the nullmodels, i.e. the models with the random interactions, in 52.8% and 62.3% of cleptoparasitic bees, 36.4% and 43.2% of oligolectic bees and 13.4% and 14.4% of polylectic bees (fig. 5; Appendix S3). When the threshold was within the 25% best performing models, these percentages increased up to 83% and 94% of the cleptoparasitic bees, 56.8% and 63.6% of the oligolectic bees and 40.2% and 51.2% of the polylectic bees. The polylectic bees showed a less skewed distribution, but more of a continuous distribution, suggesting a less selective response to the interacting species and a more general preference of biotic interactions (fig. 5A).

Flower and host specialization and interacting species’ distribution range

In general, the contribution of the interacting species to the modelled species increased as the geographic distribution of the interacting species decreased and the flower specialization increased. Both the GLMs of the flower visiting bees and the cleptoparasitic bees resulted in a set of candidate models with the body size, distribution of the interacting species and diversity of the interaction network as explanatory variables with the coefficient ($\beta$) and the contribution of the interacting species as the response variable (Appendix S7). The inverse link function was used for the GLM and the model for the flower visiting bees with the lowest AICc (AICc = 960.8) resulted in the formula with flower specialization ($\beta_1$) and distribution of the interacting species ($\beta_2$): $y = (0.063 - 0.0099 * \beta_1 + 0.013 * \beta_2)^{-1}$ ($R^2 = 0.39$; fig. 6A & 6B). The selected model of the cleptoparasitic bees contained the distribution of the host bee ($\beta$) and it had the formula: $y = (0.13 + 0.0099*\beta)^{-1}$ ($R^2 = 0.47$) (fig. 6C; Appendix S7).

Discussion

Spatial scale matters for biotic interactions in SDMs

Non-parasitic wild bees are central-place foragers that repeatedly return to their nest (Cresswell et al. 2000). Consequently, foraging habits are limited in range. Oligolectic bees nest close to their pollen plants (Gathmann & Tscharntke 2002) and the majority of resource consumption by smaller bees is within a few hundred meters of their nests (Zurbuchen et al. 2010; Hofmann et al. 2020) with larger average foraging ranges for larger-bodied bees, over 1 km (Greenleaf et al. 2007). Obligate cleptoparasitic bees lay their eggs on the pollen deposits of other bees (Litman 2019). They may search freely over large distances and are less limited in range, as they don’t have to return to their nest to collect nectar and pollen for their offspring (Litman 2019). Nevertheless, they can often be found close to the nests of their hosts, waiting for the host bee to leave and forage (Litman 2019). Our results support the recorded foraging distances and that parasites are found in proximity to their hosts, showing an optimal resolution of the interacting species as
explanatory variable at 500 m for the cleptoparasitic bees and 500 m -1 km for the oligolectic and polylectic bees). In general, the relation between the spatial resolution of the analysis and the movement of the species is most likely dependent on the type of biotic interaction.

**Genus-level biotic information as a surrogate for species-level knowledge in SDMs**

The interacting species had a higher contribution at genus taxonomic level to the modelled species at finer resolutions (500 m and 1 km). Pollinators generally visit closely related plants more often than would be expected (Vamosi et al. 2014) and our results suggest that at finer resolutions biotic interactions at genus level could adequately substitute species level interactions. This might be due to similar habitats occupied by host species and niche conservatism, observed in, for example, higher plants (Prinzing et al. 2001), or it could also imply that the genus records are dominated by the same host species at these locations. Another possibility could indicate high quality habitat for plants/hosts in general (Widhiono et al. 2016). Only a few of the oligolectic bees are monolectic, as most oligolectic bees collect pollen from more than one taxonomically related plant species (Cane 2020), resulting in a dependency on multiple plant species in the same genus. The similar biosynthetic pathways in related plants are associated with similar nutritional values of their pollen (Ruedenauer et al. 2019), which explains why plant genus is a good approximation of the biotic interaction. The cleptoparasitic bees showed lower variable importance of the interacting species at coarse resolutions at the genus level compared to the interacting species at species level. It is likely that this trend is related to a loss of information on a coarser scale, as the contribution of the interacting species at genus level decreases as the resolution decreases. A potential explanation for the higher contribution of the host at species level is that cleptoparasitic bees tend to become more specialized as the coevolution between a parasite and its host often leads to specialization (Bogusch et al. 2006). It is, therefore, no surprise that around a quarter of the European cleptoparasitic bees parasitize on only one species (Bogusch et al. 2006).

**Including biotic information in SDMs generally improves model performance**

In this study, we showed that the addition of the biotic interaction increases model performance under all metrics and this increase is the highest for the cleptoparasitic bees, followed by the oligolectic and polylectic bees. An increase in model performance by the addition of the host of the cleptoparasitic bee has been observed (Gianinni et al. 2013), however the increase in model performance for oligolectic bees and their host plants at a resolution of 10 km was often not significant (Giannini et al. 2013). This highlights once again the importance of resolution. Furthermore, the importance of including biotic interactions is not limited to plant-pollinator interactions (Heikkinen et al. 2007; Kissling et al. 2007; Bateman et al. 2012; Leach et al 2016; Mpakairi et al. 2017; Roslin et al. 2017; Atauchi et al. 2018; Herrera et al. 2018; Mathieu-Bégué et al. 2021) and biotic interaction can play a role in the distribution range edges of species even at a larger scale (Paquette & Hargreaves 2021; Freeman et al. 2022).

The null models with random pollinated plants or host bees, revealed how the specificity of the interaction (e.g. specialist versus generalist) influences which interacting species could be used. A higher specificity was observed for the oligolectic and cleptoparasitic bees than the polylectic bees, whose models benefitted from a range of different flowering plants. The high performance of the specific interacting species in the models of the cleptoparasitic bees confirmed how important their host species are for modelling their distribution. Another contributing factor can be the biases in the data source: the distribution of the cleptoparasitic interacting species is sourced from the same wild bee occurrences database and therefore, shares similar collection biases to the modelled species. In contrast to the distribution of the plant species which likely have their own separate collection biases. In the case that data on the interacting species is lacking, an option would be to use information from a co-occurring species (Briscoe Runquist et al. 2021). Our study showed that the inclusion of other visited plants can also improve model performance.

We found that a higher degree of flower specialization for the flower visiting bees and a narrower distribution of the interacting species for both the cleptoparasitic and flower visiting bees were related to a higher importance of the biotic factor in the SDMs. The dependence between the distribution of two organisms, each at one side of the biotic interaction, has been shown in many different studies (Fauchald et al. 2000,
Byholm et al. 2012, Atauchi et al. 2018) and even at a macroecological scale (Araújo & Luoto 2007). If
the distribution of the interacting species is narrow, it is more likely to be a limiting factor, delimiting
the boundaries of the potential distribution of the modelled species. Other studies have found that narrow
distributions, although not too narrow, mean more accurate models and high importance for certain key
habitat factors (Tsoar et al. 2007; Syphard & Franklin 2010) and that specialist species yield better models
than generalist species (Marshall et al., 2015; Grenouillet et al. 2011). We found that if the biotic relationship
is strong, e.g. parasite-host relationships, then that becomes by far the most important factor.

**Considerations on how to optimally include biotic information in SDMs**

Methodological considerations include whether the biotic interaction should be included as the raw distri-
bution or the predicted suitability map and whether a classical modelling approach should be used or a
Joint species distribution modelling (JSDM) approach. When the choice is made to introduce the predicted
suitability maps of the interacting species’ instead of the raw distributions, results should be interpreted
with care, as the prediction map may be highly dependent on abiotic factors (such as climatic variables),
making it less likely to explain the species occurrences which abiotic factors cannot explain (Silva et al.
2014). Another commonly used approach for modelling biotic interactions is JSDM and it is suitable for
situations where the biotic interactions are not known a priori, and this method helps to understand a
species’ geographical range from a community ecology perspective (Pollock et al. 2014; Ovaskainen et al.
2017). The risk is, however, that any detected relationships between species may be due to shared habitat
preferences not accounted for elsewhere in the model instead of biotic relationships (Wisz et al. 2013; Pollock

Here, we showed that biotic factors can improve the SDMs of wild bees in the Netherlands, especially when
the distribution of the interacting species is narrow. Resolution, taxonomic level, the degree of specificity
in the interaction, e.g. specialist species vs. generalist species, should be taken into account to achieve the
most optimal models. We recommend using single species or genus data as a biotic variable in the models of
specialist species and to use an approximate, such as flower richness, for more generalist species.

**Acknowledgements:**

We would like to thank the European Invertebrate Survey in the Netherlands for providing data on wild
bees in the Netherlands. This study was supported by funds from the Multisectoral Analysis of climate and
land use change impacts on Pollinators, Plant diversity and crop Yields project (MAPPY; SC5-30-2017) and
Netherlands’ Science foundation (NWO).

**Figure captions:**

**Figure 1:** Schematic overview of the modelling workflow. The elements are represented in boxes and consist
of models, information from databases or variables (brown), variables, indices, evaluation measures or ex-
perimental conditions (e.g. no data is used or the resolution of the used variables). The arrows indicate the
information flow. The evaluation measures are: (i) the corrected Akaike information criteria (AICc; Burnham
& Anderson 2002), (ii) the continuous Boyce index (Boyce index; Hirzel et al. 2006) and (iii) area under the
curve of the receiver operating characteristics curve (AUC) applied to species distribution models (Elith et
al. 2006).

**Figure 2:** The effect of the resolution and taxonomic level on the contribution of the biotic variable to
the model, expressed as the ranking of the biotic variable contribution per species (from high to low: 1-5;
2A) and the difference in variable contribution between the species that the modelled species interacts with
(interacting species) added at species and genus taxonomic level per species (2B). The arrows indicate the
direction, where the variable contribution is the highest for the respective taxonomic level. The resolution
is the scale in longitudinal and latitudinal direction at which the interacting species is observed. The gray
area is the standard deviation.

**Figure 3:** The differences between models including host plant or parasitic host interactions and models
with only land use, climate and soil variables. Evaluation measurements include Area Under the receiver
operating characteristic Curve (AUC) value of the evaluation dataset (figure 3A), Continuous Boyce Index (CBI) of the evaluation dataset (figure 3B), Aikake Information Criteria for small sample sizes for both the evaluation and calibration data (AICc; Burnham & Anderson 2002; figure 3C) and AICc of the evaluation data only (figure 3D). Host plants and hosts of parasites were either included at the species or genus level. The difference in evaluation metrics for models with and without biotic factors, or difference from zero, is tested for significance with a One-Sample Wilcoxon Signed Rank Test ($p<0.05$; $p<0.01$, $p<0.001$). For the AICc both the calibration and evaluation dataset were included, because 66 modelled bee species did not have enough evaluation datapoints to calculate the AICc.

**Figure 4:** The different boxplots represent the summed contribution of the five climate variables, the sixteen land use variables, the eight soil variables and the single biotic variable, averaged over the modelled species in the functional groups. The biotic variable is averaged over the species and genus taxonomic level of the visited plant or host bee. The different letters above the boxplots indicate significant differences between variable groups within the functional trait groups ($p<0.05$).

**Figure 5:** The comparison of the biotic interaction models to models with random interactions, described as null models, with plants (for the oligolectic and polylectic bees) or bees (for the cleptoparasitic bees). Figure 4A shows the distribution of the performance of the biotic interaction models, expressed as the rank of the evaluation AUC among all interaction models divided by the total number of models. The y-axis represents the total number of modelled species that fall within the performance threshold on the x-axis. For example, the performance in evaluation AUC of the known interaction was compared to the other 306 plant species and ranked based on the position. If the known interaction was the third best performing model, the modelled species would have the value of 0.98% (the percentage rank would be $3/307 * 100 = 0.98\%$) and fall within 0-2.5% best performing models. The two lines indicate the threshold of 5% and 25% best performing models. Figure 4B summarizes the results, comparing the percentage of modelled species that fall within the 5% best performing ranks, indicating a significant difference from the null models with $p < 0.05$ (5% best performing models), and 25% best performing ranks. Although the percentage of models that fall within the 5% best performing models is higher for the oligolectic bees and cleptoparasitic bees, the polylectic bees show a high percentage of performance within the 25% best performing models, showing a more general preference of biotic interactions. The number of random interactions for every null model are 306 interactions for the flower visiting bees with the interacting species at species level, 160 interactions for the flower visiting bees with the interacting species at genus level, 99 interactions for the cleptoparasitic bees with the interacting species at species level and 100 or 15 interactions for the cleptoparasitic bees with the interacting species at genus level (see Appendix S3 in supporting information).

**Figure 6:** The results of the Generalized Linear Models (GLMs) show the relation between flower specialization (Shannon-Wiener index of number of plants genera interacted with) and the contribution of the biotic variable to the models of the oligolectic and polylectic bees (Shannon 1948; figure 6A). Figure 6B shows effect of distribution of the most visited genus on the contribution of the biotic variable to the model. Figure 6C shows the relation between the distribution of the host bees and the contribution of the biotic variable to the models of the cleptoparasitic bees.

**Supporting information:**

**Appendix S1:** Additional information materials and methods

**Appendix S2:** Model optimization: regularization multiplier

**Appendix S3:** Details null models

**Appendix S4:** Detailed results resolution and taxonomic level

**Appendix S5:** Detailed results evaluation measures

**Appendix S6:** Comparing the contribution of the interacting species with climate, land use and soil variables

**Appendix S7:** Details of the generalized linear models

**Appendix S8:** References of supporting information

**Table S1:** Overview of the different land use classes used.

**Table S2:** Overview of the different soil categories used.

**Table S3:** Comparison of plant pollen dependencies and most visited plants.

**Table S4:** Cleptoparasitic bees and their hosts used from the literature (Peeters et al. 2012).

**Table S5:** The standardized ODMAP protocol for reporting species distribution models from Zurell et al. 2020.
References:
