Chapter 3:

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

Based on: Moens, M., Biesmeijer, J. C., Huang, E., Vereecken, N. J., & Marshall, L. (2024). The importance of biotic interactions in distribution models of wild bees depends on the type of ecological relations, spatial scale and range. *Oikos*, e10578.

Abstract

1. Studies have found that biotic information can play an important role in shaping the distribution of species even at large scales. However, results from species distribution models are not always consistent among studies and the underlying factors that influence the importance of biotic information to distribution models, are unclear.
2. We studied wild bees and plants, and cleptoparasite bees and their hosts in the Netherlands to evaluate how the inclusion of their biotic interactions affects the performance of species distribution models. We assessed model performance through spatial block cross-validation and by comparing models with interactions to models where the interacting species was randomized. Finally, we evaluated how, (i) spatial resolution, (ii) taxonomic rank (genus or species), (iii) degree of specialization, (iv) distribution of the biotic factor, (v) bee body size and (vi) type of biotic interaction, affect the importance of biotic interactions in shaping the distribution of wild bee species using generalized linear models.
3. We found that the models of wild bees improved when the biotic factor was included. The model performance improved the most for parasitic bees. Spatial resolution, taxonomic rank, distribution range of the biotic factor and degree of specialization of the modelled species all influenced the importance of the biotic interaction to the models.
4. We encourage researchers to include biotic interactions in species distribution models, especially for specialized species and when the biotic factor has a limited distribution range. However, before adding the biotic factor we suggest considering different spatial resolutions and taxonomic ranks of the biotic factor. We recommend using single species or genus data as a biotic factor in the models of specialist species and for the generalist species, we recommend using an approximate measure of interactions, such as flower richness.

3.1 Introduction

During the last decades, there has been a massive increase in the number of published studies using species distribution models (SDMs), reviewed in (Lobo et al. 2010, 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

and Leathwick 2009), including: the estimation of potential distributions under different climate change scenarios (Marshall et 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).
Biotic information is rarely included directly in distribution models. Biotic interactions are interactive, resource-related variables and are known as bionomic variables, relating to the Eltonian niche (Soberón 2007). On the other hand, the environmental variables represent the non-interactive scenopoetic variables and are related to the Grinnellian niche (Soberón 2007). Even though competition can drastically change the distribution of species at a small scale, if the scale is large enough competitors may coexist for long times and the effect of the bionomic variable may be less apparent (Soberón 2007). The mathematical background of distribution modelling assumes that the predictor variables are independent from the modelled species and such unlinked variables are defined as scenopoetic variables (Anderson 2017). Biotic factors were not considered as scenopoetic variables, as the biotic predictor variable is influenced by the modelled species that they interact with (Anderson 2017). However, a biotic factor that is affected minimally by the modelled species, would constitute a scenopoetic variable and therefore the biotic factor could be a valid and valuable addition to the models (Anderson 2017). 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-Bégné et al. 2021), facilitative (Heikkinen 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 (Araújo and Luoto 2007, Espíndola and Pliscoff 2019, Kass et al. 2020).

The extent to which the inclusion of a biotic factor will improve the accuracy of a SDM depends on several properties of the model and the characteristics of the ecological interaction. For example, the spatial scale of the biotic factor is of importance when introducing it as an explanatory variable (Heikkinen et al. 2007). 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 and spatial resolution may strongly affect the contribution of biotic interactions to modelled distribution patterns (Pearson and Dawson 2003, Soberon and Peterson 2005, Wisz et al. 2013). Heikkinen et al. (2007) showed that the impacts of facilitation between owls and woodpeckers are more visible in terms of model performance at a resolution of 10 km than 40 km. This is consistent with Pearson & Dawson (2003), who hypothesized that at broader scales and coarse resolutions, climate variables are more dominant and biotic interactions less apparent. However, the type and strength of an ecological interaction may influence the scale dependency. For example, an obligate parasite with a strong interaction with its host may always be more dependent on its hosts' distribution at any resolution.
Additional distinguishing attributes of the biotic factor include the taxonomic rank of the biotic factor, a crucial factor for specialist bees and their preferred plant taxon (Robertson 1925, Rasmussen et al. 2020). Characteristics of the modelled species themselves are also important, such as the degree of specialization (e.g. generalists vs. specialists) and movement range of species, which is strongly linked to how much the limited distribution range of the biotic factor may restrict the distribution of the modelled species (Giannini et al. 2013). 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. The dietary breadth of the species could influence the importance of the biotic factor in the models (e.g. specialist vs generalist; de Araújo et al. 2014). In the case of bees, it has been shown that the population trend of specialist bees is correlated to the population trend of plants that they are dependent on for their pollen (Scheper et al. 2014b). Specialist bee species have a tendency to decline more than generalist bee species and their decline is correlated to the population trend of their host plant (Biesmeijer et al. 2006) and this leads us to expect that the specialist species show a higher contribution of the biotic factor to their models. Not much is known about the effect of the distribution range of a biotic factor on its importance to SDMs. However, it is expected that a smaller distribution range of the biotic factor would have a higher contribution to the models of specialist species, as it more likely to be the limiting factor of the modelled species.

Here, we aim to use *a priori* knowledge to investigate the factors that influence the importance of biotic interactions in species distribution models of wild bees. 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) (Robertson 1925, 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 suitable 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 performance of biotic factors in explaining distributions of wild bees in the Netherlands and (2) assess how different factors influence the importance of the biotic variable, more precisely (i) spatial resolution, (ii) taxonomic rank, (iii) degree of specialization, (iv) distribution of the biotic factor, (v) bee body size (as a proxy for movement range) and (vi) type of biotic interaction (cleptoparasitic, oligolectic and polylectic bees).

3.2 Materials and Methods

Overview of the modelling approach

The methods in this paper can be subdivided into three separate modelling approaches (fig. 1). For the first aim of the study, assessing the performance of biotic factors in explaining distributions of wild bees in the Netherlands, we developed SDMs with the modelled bees as focal species and the abiotic variables and the biotic factor as predictors. The biotic factor was either the most visited plant for the pollen-collecting bees or their known host bee for the cleptoparasitic bees. The species of the biotic factor was determined using the literature or a database and therefore these models are referred to as known interaction species distribution models (KI-SDMs). In contrast to the known interactions, we also introduced a random species as the biotic factor; these models are referred to as randomized interaction species distribution models (RI-SDMs). These models are an adapted version of the SDM null models (Raes and ter Steege 2007) and they account for collection biases and can test whether the importance of the biotic factor is specific to the known interaction or might be accounted for by the interaction with any other plant or bee. The second aim of this study is to assess how different factors influence the importance of the biotic factor to the models. To compare the effects of (i) spatial resolution and (ii) taxonomic ranks we extend the KI-SDMs with a set of extra models that vary in the spatial resolution and taxonomic rank of the biotic factor. The third modelling step utilizes the output from the SDMs (the variable importance of the biotic factor) to assess the relative importance of the other factors, namely (iii) degree of specialization, (iv) distribution of the biotic factor and (v) bee body size (which relates to movement range) in explaining the contribution of the biotic factor to the models. We approached this by fitting generalized linear models (GLMs) with variable importance as the response variable and with the above factors included as explanatory variables.

Abiotic variables

The abiotic variables totalled 29 variables, including five climate variables, sixteen land use variables and eight soil variables. We used climate data from the Koninklijk Nederlands Meteorologisch Instituut (KNMI 2016) from the period of 2000 to 2015 and converted the temperature and precipitation values to the standard 19 bioclimatic variables (Fick and Hijmans 2017) using the R package dismo (version 1.3-3) (Hijmans et al. 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 data was 100 m (100 m by 100 m). Land use data consisted of 15 land use categories from different vector shapefile sources (see Appendix table 1 in the electronic supplementary material; Centraal Bureau voor de Statistiek (CBS), 2012; Inter Provinciaal Overleg, 2016; Ministerie van Economische Zaken, 2015). As an additional variable the sum of the different number of land use polygons was calculated per 100 m x 100 m grid cells by summing the different number of land use types, in the corresponding grid cells in the raster package (version 3.6-3) in R (Hijmans 2018). This package was also used to rasterize the land use and soil shapefiles to percentage cover per 100 m x 100 m grid cell. The soil data consisted of 8 classes of soil types in the Netherlands representing different concentrations of sand, silt and clay (see Appendix table 2 in the electronic supplementary material; Grondsoortenkaart, 2006). The climate, land use, and soil data were not strongly correlated (Spearman's $\rho < 0.7$; Dormann et al., 2013).

Biotic variables

The bee occurrence data used in this study consisted of (i) opportunistic observations of bees from 2004 to 2019 and (ii) bee-flower visitation records from 2004 to 2019, obtained from the European Invertebrate Survey Netherlands (EIS 2020). Both the occurrences and bee-flower visitation records originate from the same database that is a compilation of different datasets with observation records collected by professionals, amateur experts and citizen scientists. All the data has been validated by professionals. The different datasets consist of both opportunistic data and structured surveys. For the bee occurrences, we assigned a value of one when at least one occurrence point was present in a grid cell of 100 m x 100 m and a value of zero when no occurrence point was present in the cell. Bees were classified as bees that visit a single plant taxon or show a clear preference for a single plant family or genera (oligolectic), bees that collect pollen and nectar from various plant taxa (polylectic) (Robertson 1925, Rasmussen et al. 2020) and brood parasites (cleptoparasitic); these traits were based on a traits database created for the Status and Trends of European

Pollinators (STEP) project (Potts et al. 2015). We discarded the species without flower visitation information and those with fewer than fifteen observations This selection resulted in 44 oligolectic bees, 97 polylectic bees and 55 cleptoparasitic bees, totalling 194 species of the more than 300 bee species in the Netherlands. We determined the most visited plant species and genus for each pollen-collecting bee species using the bee-flower visitation database. These distributions of these plant species were then used as the biotic factor in the KI-SDMs for the pollen-collecting bees. We checked the 44 oligolectic bees for their most visited plant from the flower visitation database and found that in all cases that the plant genus or species they had been observed visiting the most was also listed as their pollen-collecting plant taxon in the traits database (Appendix table 3 in the electronic supplementary material). Since no quantitative interaction data was available for cleptoparasitic bees and their host, we applied the knowledge from the literature to determine all the known host bee species and genera for the cleptoparasitic bees (Appendix table 4 in the electronic supplementary material; Peeters et al., 2012).
The plant occurrences for the period 2004-2019 were obtained from Dutch National Database of Flora and Fauna (NDFF 2021). The plant occurrence database like the bee occurrence database includes data from different sources. The data is either opportunistic data or structured vegetation surveys both from professionals and amateurs. As with the bee occurrence data, only entries that are validated by professionals are included. The occurrences of the biotic factor, both host bees and plants, were transformed to a raster at a resolution of 100 m. We assigned a value of one when at least one occurrence point was present in the cell and a value of zero when no occurrence point was present in the cell. We repeated this process to create biotic factors to include in the KI-SDM at resolutions of 500 m, 1 km, 5 km And 10 km by giving a positive value to a cell, when an occurrence point would be present in a larger aggregated grid cell of the respective distances. The same procedure was also applied at the taxonomic rank of genus. Some of the cleptoparasitic bees had multiple hosts and in this case a grid cell was classified as presence if any of the host species or genera was present. The models were run with the other variables at a 100 m resolution to eliminate any effects that their aggregation might have. All the biotic factors were tested for correlation with the abiotic variables and only 2.42% had a significant correlation that was higher than 0.7 Spearman’s rho (see Appendix data 1 in the electronic supplementary material).

Known interaction species distribution models

For the modelling of the species distributions, we executed 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. 2017) and each model was divided in four spatially distinct selections of training and testing datasets, using the spatial block validation method in the ENMeval package version 2.0.3 (Kass et al. 2021). Model evaluation measures were always averaged across these four spatial folds. The study area, where 10,000 background points were drawn from, included the administrative borders of the Netherlands, excluding grid cells with only sea. An overview of the SDMs can be found in fig. 1 and Appendix table 5 in the electronic supplementary material, following the ODMAP protocol for SDM metadata reporting (Zurell et al. 2020). We ran all models in parallel using the package snow version 0.4-3 (Tierney et al. 2008) and the package parallel in base R version 4.0.3 (R Core Team, 2013). Model performance was assessed with the average area under the curve of the receiver operating characteristic (AUC). This evaluation measure is a threshold independent evaluation method of the performance of the calibration and evaluation dataset (Elith et al., 2006; Phillips et al., 2006). This metric has been criticised for its use with presence only SDMs because of its sensitivity to imbalanced data and requirements for absence data (Lobo et al. 2008), and for this reason we included the corrected Akaike information criterion (AICc; Burnham & Anderson, 2004), and Continuous Boyce Index (CBI) (Hirzel et al. 2006) as additional evaluation measures. Unlike the AUC the AICc includes a penalty based on the number of variables that is used in the models. We also looked at the percentage contribution of the variable to the model gain and the permutation importance, which are measures of variable importance calculated in the MaxEnt algorithm (Phillips et al. 2006). In the converged model, all variables are kept constant, while one is changed among the presence and background points and the corresponding change in AUC is used for the calculation of the permutation importance (Jarnevich et al. 2016). The percentage contribution can vary per environmental variable, as the MaxEnt algorithm can take different paths to come to the final model, in contrast to the permutation importance, that remains constant (Phillips et al. 2006). The degree of overfitting was used to select the appropriate regularization multiplier among the values of 1, 3, 5, 7, 9, 11, 13 and 15. We chose a value of 5 based on the evaluation AUC – calibration AUC, an indication of overfitting (Radosavljevic and Anderson 2014; Appendix text 1 in the electronic supplementary material) and the number of features and evaluation AUC. We selected only simple features (linear and quadratic features) in the model settings instead of more complex feature types, that would complicate the ecological interpretation of the models' responses to (a)biotic variables (Syfert et al. 2013). Additionally, allowing more complex features may result in a response that results from the database structure or data collection instead of representing physiological and/or ecological relationships with the environment (Syfert et al. 2013). The percentage contribution and permutation importance of the biotic factor 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. By ranking the models, differences in variable contribution between species are removed. We made prediction maps of the models with the biotic factor (plant species for pollen-collecting bees and bee host for cleptoparasitic bees) at a resolution of 1 km for the calculation of the models of the four spatial folds of the spatial block validation were used to calculate the evaluation measures and the average of these four values was taken. Per modelled species, this resulted in eleven models: one model without the biotic interaction, five models with the biotic interaction at genus taxonomic level at five different resolutions and five models with the biotic interaction at species taxonomic level at five different resolutions. Presence and absence maps were made based on the maximum training sensitivity and specificity threshold, that integrates both measures for the estimation of prediction success (De Barros et al. 2012). The difference between the CBI, AICc and evaluation AUC of models with and without biotic factors was not normally distributed and it was tested for significance against the null hypothesis of no significant difference by using a one-sample Wilcoxon signed rank (Wilcoxon 1947). We also compared the percentage contribution of the aggregated variables classes (land use, climate, soil, and biotic factors), using a non-parametric Kruskal-Wallis H test (Kruskal and Wallis 1952) with a post-hoc Nemenyi test (Sachs L 1997) for pairwise group comparisons in the PCMCRplus package (version 1.9.7) (Pohlert 2023).

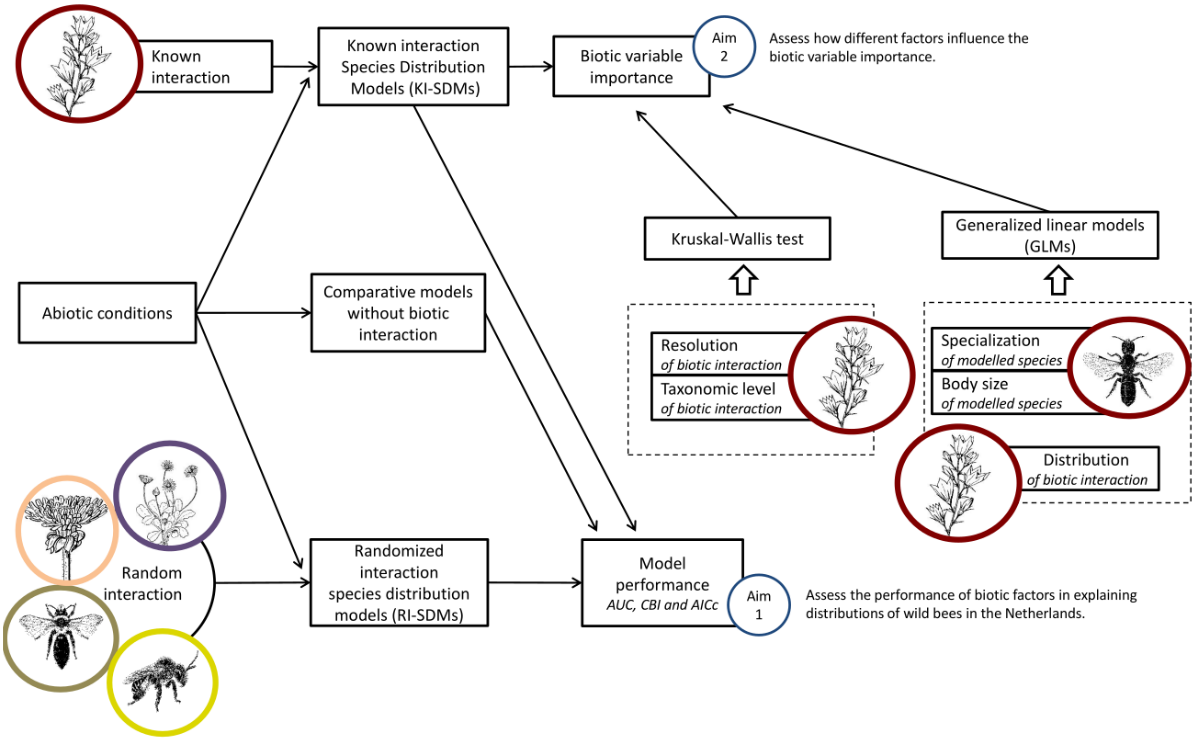
Known interaction
Known interaction Species Distribution Models (KI-SDMs)
Biotic variable importance
Aim 2
Assess how different factors influence the biotic variable importance.
Kruskal-Wallis test
Generalized linear models (GLMs)
Abiotic conditions
Comparative models without biotic interaction
Resolution of biotic interaction
Taxonomic level of biotic interaction
Specialization of modelled species
Body size of modelled species
Distribution of biotic interaction
Random interaction
Randomized interaction species distribution models (RI-SDMs)
Model performance AUC, CBI and AICc
Aim 1
Assess the performance of biotic factors in explaining distributions of wild bees in the Netherlands.

Figure 1: Schematic overview of the modelling workflow. The boxes represent models and variables and the arrows indicate the information flow. The circles represent the research aims that target the respective evaluation measures.

Randomized interaction species distribution models

The RI-SDMs differ from the KI-SDMs in the species that is included as the biotic factor. These RI-SDMs address the importance of including biotic interactions by comparing the KI-SDMs to multiple models which are structurally equivalent except that a different (random) species is selected as the biotic factor. The biotic factor consists of plant species (for pollen-collecting bees) or host bee species (for cleptoparasitic bees) that are randomized from a pool of all plants that are visited by bees (interaction database; EIS 2020) and all host bees of cleptoparasitic bees (from the literature; Peeters et al., 2012). These RI-SDMs are then compared to the KI-SDMs and allow us to distinguish the specificity of the known interaction to the models (fig. 1). For example, a model may benefit from many different plant or bee species as opposed to a model that benefits from only a single species and that may represent a more specific interaction. All the potential visited flowers were included in the total number of RI-SDMs per pollen-collecting bee (307 plant species and 161 plant genera). The cleptoparasitic bees had multiple hosts and for the RI-SDMs we used a total of 100 models per modelled species, randomizing the multiple hosts from a set of 15 potential host bee genera or 42 host species. A more detailed description of the RI-SDMs can be found in Appendix text 2 in the electronic supplementary material.

The evaluation AUC of the KI-SDM was ranked among the RI-SDMs per focal species and the percentage rank of the known interaction amongst the randomized interactions was compared between groups. We analysed per species the ranking of the model with the biotic factor 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

The GLMs address the third research question in this study, assessing the importance of flower and host specialization, distribution range of the biotic factor and bee body size (which relates to movement range) in explaining the contribution of the biotic factor to the KI-SDMs. These variables are included as explanatory variable in the GLMs with the variable importance of the biotic factor to the KI-SDMs as the response variable (fig. 1). To calculate the distribution range of the biotic factor for the flower visiting bees, we summed the amount of grid cells occupied by each visited plant species from the plant observation data. 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 inverse of the Shannon-Wiener index (Shannon 1948). As oligolectic bees may visit multiple species of the same genus, we decided to calculate the diversity of the genera visited and not the species themselves. Thirdly, we used the information on body size from the bee trait database. Intertegular distance (ITD, in mm; the distance between the wing insertion points) was used as a proxy for body size (Greenleaf et al. 2007). The cleptoparasitic bees were modelled in a similar way, except that the distribution range of the biotic factor was calculated from observation data from potential host bees (Appendix table 2 in the electronic supplementary material) and the host specialization was the number of potential hosts in the literature (Peeters et al. 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 and Bravington 2013). Model selection was performed using the AICc and if the difference between models was less than two AICc units, we selected the models with the fewest variables and with only significant coefficients. From the focal species in the KI-SDMs, we made a selection of those species for the GLMs, removing the species that did not find a contribution (e.g. no features of the respective variable present in the model) of their biotic factor in the KI-SDMs. The three explanatory variables of body size (a), distribution range (b) and specialization (c) resulted in seven possible combinations of variables: a+b+c, a+b, a+c, b+c, a, b, c (Appendix text 3 in the electronic supplementary material). 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, 2013).

3.3 Results

The effect of the known interaction on overall model performance

The inclusion of biotic factors resulted in a statistically clear improvement of the KI-SDMs in explaining distributions of wild bees in the Netherlands. This was true for all evaluation measurements considered; Area Under the Curve (AUC), continuous Boyce index (CBI) and corrected Aikake information criteria (AICc). The final KI-SDMs included the known interaction at a resolution of 1 km, which was on average the optimal resolution for most models (see section “Influence of spatial resolution and taxonomic rank on model performance”). The models of the cleptoparasitic, oligolectic and polylectic bees all showed a statistically significant increase in evaluation AUC (fig. 2A), evaluation CBI (fig. 2B) and a decrease in calibration and evaluation AICc (fig. 2C; Appendix text 4 in the electronic supplementary material). Even though a similar trend was visible for the evaluation AICc, 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 at the species and genus level (fig. 2D). The biotic interaction had a high contribution to the KI-SDMs of all the pollen-collecting and cleptoparasitic bees relative to the climate, land use and soil variables (fig. 3; Appendix text 5 in the electronic supplementary material). For the cleptoparasitic bees specifically, the biotic interaction always had the highest contribution to the KI-SDMs.

When comparing between the pollen-collecting bees and cleptoparasites , the models of the cleptoparasitic bees showed a statistically significant greater improvement in evaluation AUC, evaluation CBI and in calibration and evaluation AICc compared to the oligolectic and polylectic bees at both taxonomic ranks (Appendix text 4 in the electronic supplementary material). The difference in evaluation metrics between oligolectic and polylectic bees was in no case significant.

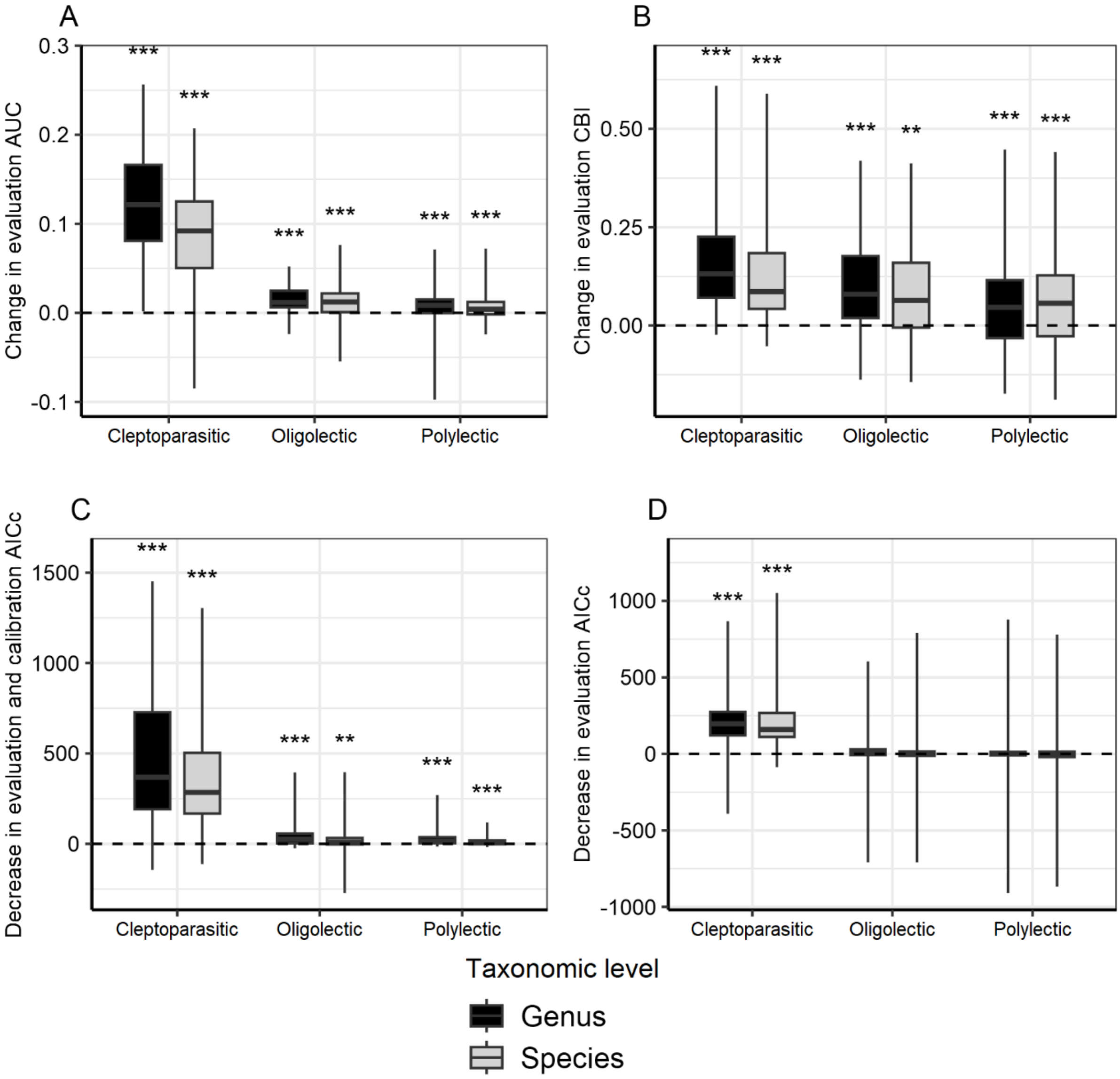

Figure 2: The differences between models including host plant or parasitic host interactions and

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 (fig. 3A), Continuous Boyce Index (CBI) of the evaluation dataset (fig. 3B), Aikake Information Criteria for small sample sizes for both the evaluation and calibration data (AICc; Burnham & Anderson 2002; fig. 3C) and AICc of the evaluation data only (fig. 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.

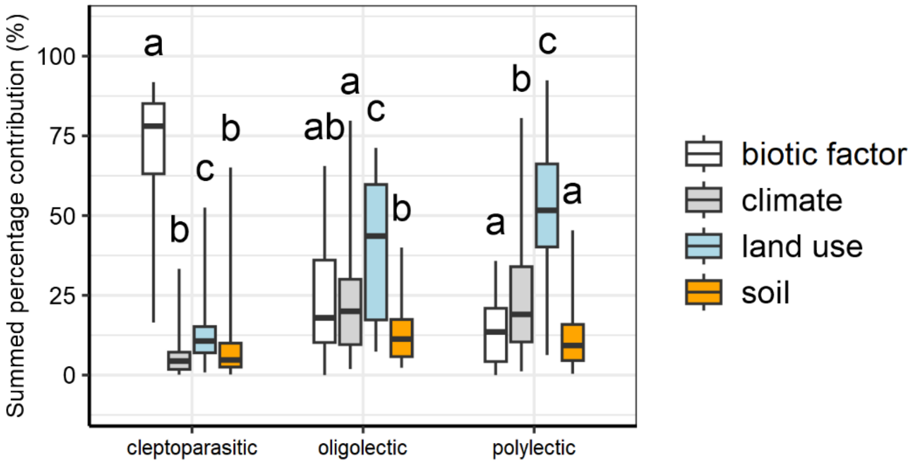

Figure 3: The different boxplots represent the summed percentage contribution of the five climate variables, the sixteen land use variables, the eight soil variables and the single biotic factor, averaged over the modelled species in the groups. The biotic factor is averaged over the species and genus taxonomic rank 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).

The effect of any interaction on overall model performance

The purpose of the RI-SDMs was to check whether the biotic factor only improved the models, when it was included as an, ecologically supported, known interaction of a species instead of a randomized interaction. A higher specificity to the known interaction would provide a stronger support for the inclusion of known biotic interactions into SDMs. Models with the biotic factor (added at both taxonomic ranks) had an evaluation AUC that scored within the highest 5% best performing models of the RI-SDMs in 52.8% (species) and 62.3% (genus) of cleptoparasitic bees, 36.4% and 43.2% of oligolectic bees and 13.4% and 14.4% of polylectic bees (fig. 4B; Appendix text 2 in the electronic supplementary material). These percentage show a significant deviation in all groups with the biotic factor as the known interaction scoring higher than any interaction (exceeding the 5% threshold). When the threshold was increased to 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 a more continuous distribution (fig 4A), suggesting that the models of the polylectic bees benefit more from any interaction as opposed to the known interaction compared to the oligolectic bees.

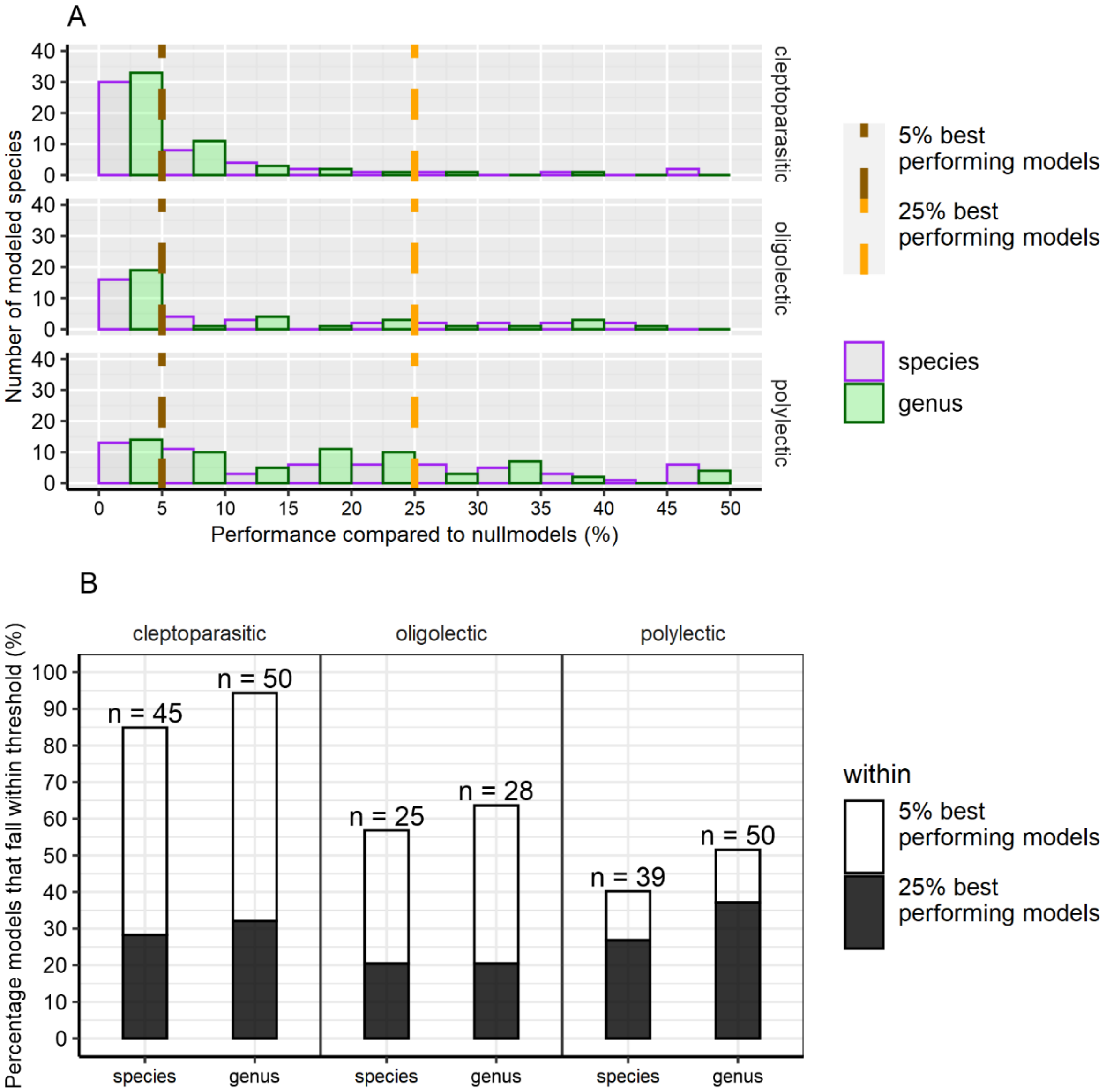

Figure 4: The comparison of the biotic interaction models to models with random interactions, described as randomized interaction species distribution models (RI-SDMs), with plants (for the oligolectic and polylectic bees) or bees (for the cleptoparasitic bees). Fig. 4A shows the distribution of the 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 focal 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. Fig 4B summarizes the results, comparing the percentage of modelled species that fall within the 5% best performing ranks, indicating a significant difference from the RI-SDMs 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 set of RI-SDMs are 306 interactions for the flower visiting bees with the biotic factor at species level, 160 interactions for the flower visiting bees with the biotic factor at genus level,

99 interactions for the cleptoparasitic bees with the biotic factor at species level and 100 or 15 interactions for the cleptoparasitic bees with the biotic factor at genus level (see Appendix text 2 in the electronic supplementary material).

Influence of spatial resolution and taxonomic rank

We found that varying the spatial resolution and taxonomic rank of the biotic factor influenced the model performance of the KI-SDMs. The percentage contribution of the biotic factor to the models was ranked highest at 500 m resolution (cleptoparasitic bees) and 1 km (oligolectic and polylectic bees) and at both species and genus taxonomic ranks (fig. 5A; Appendix text 6 in the electronic supplementary material). The optimal resolution of 500 m for the cleptoparasitic and 1 km for the polylectic bees had significantly a higher rank of the importance of the biotic factor ($p<0.05$) compared to the other resolutions except for the difference between 500 m and 1 km, which was not significant at both taxonomic ranks (Appendix text 6 in the electronic supplementary material). The oligolectic bees only showed a statistically significant difference between the optimal resolution of 1 km and the extremes (100 m and 10 km). The permutation importance showed a similar trend as the percentage contribution over the different resolutions for the oligolectic and polylectic bees and the models at a resolution between 1 km and 5 km ranked the highest (Appendix figure 1 in the electronic supplementary material). The permutation importance of the cleptoparasitic bees showed a different trend: the variable contribution ranked the highest at 100 m and 5 km resolution, when the biotic factor was introduced at the species level. When the biotic factor was introduced at genus level, it ranked the highest at 1 km, but this ranking was very close to 500 m and 100 m.

When we analysed the interaction between spatial resolution and taxonomic rank, we found that the percentage contribution and permutation importance of the biotic factor to the models was generally higher when added at the species level for coarser resolutions and at the genus level at finer resolutions. The contribution of the biotic factor 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 for both percentage contribution and permutation importance (fig. 5B; Appendix text 6 in the electronic supplementary material; Appendix figure 1 in the electronic supplementary material). This trend was the strongest for the cleptoparasitic bees with a higher percentage contribution for the biotic factor at genus taxonomic rank compared to the species taxonomic rank at 100 m, 500 m and 1 km (18.27 %, 7.94% and 2.42%) and lower percentage contribution at 5 km and 10 km (-21.00% and -22.33%).

The percentage contribution of the biotic factor was higher for the cleptoparasitic bees compared to the oligolectic and polylectic bees for both the genus and species taxonomic rank of the biotic factor. This difference was statistically supported at all resolutions of the biotic factor (Appendix text 6 in the electronic supplementary material; Appendix data 2 in the electronic supplementary material). The oligolectic and polylectic bees only showed a statistically significant difference at a 10 km resolution with a higher contribution of the biotic factor for the oligolectic bees. Similar to the percentage contribution, the permutation importance was significantly higher for the cleptoparasitic bees compared to the pollen-collecting bees for all taxonomic ranks and resolutions of the biotic factor (Appendix data 2 in the electronic supplementary material). Additionally, the oligolectic bees had a higher permutation importance compared to the polylectic bees for all taxonomic ranks and resolutions of the biotic factor (Appendix data 2 in the electronic supplementary material).

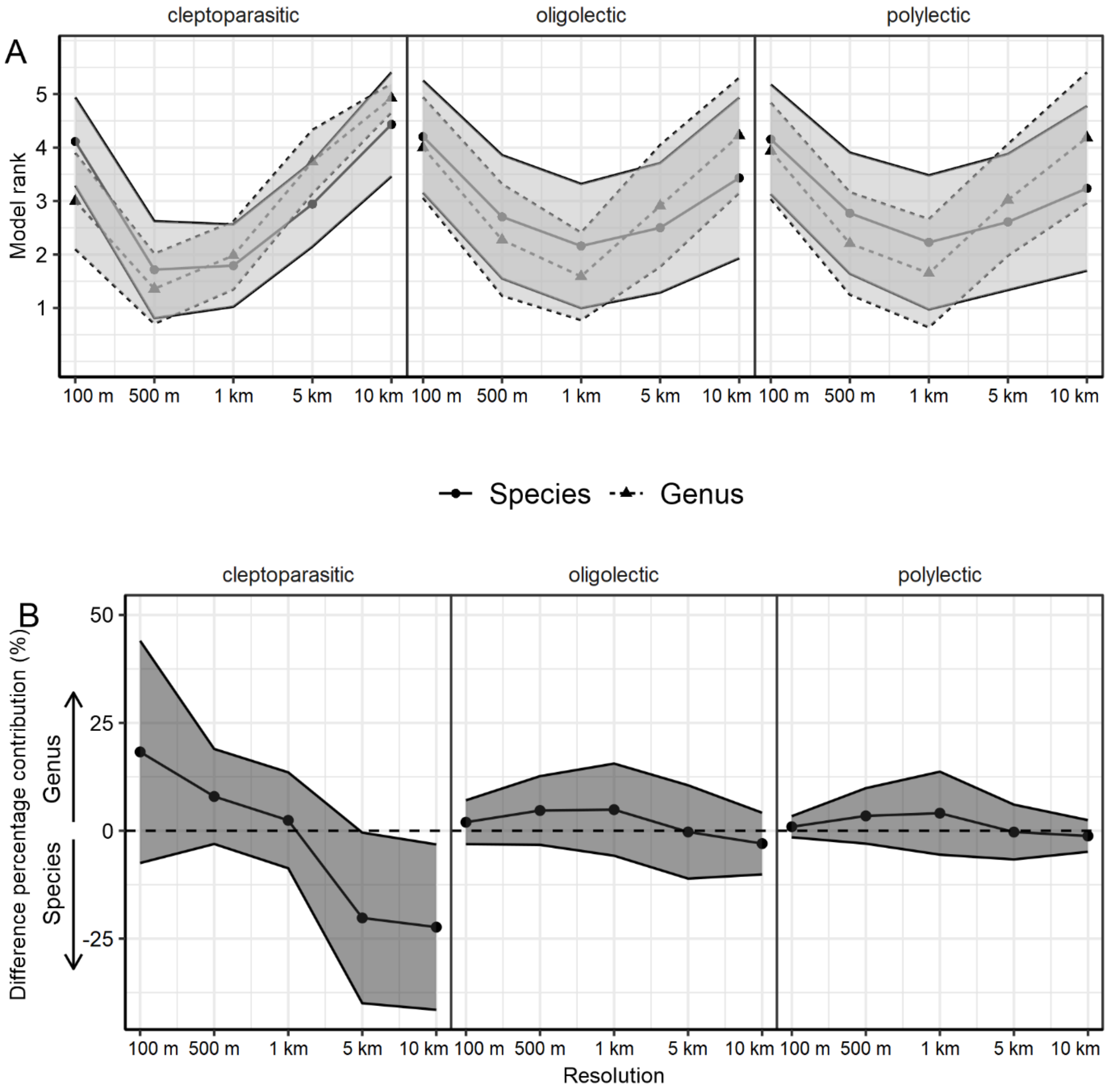

Figure 5: The effect of the resolution and taxonomic rank on the percentage contribution of the biotic factor to the model, expressed as the ranking of the biotic factor contribution per species and indicated with lines (from high to low: 1-5; fig. 5A) and the difference in variable contribution between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

Influence of flower and host specialization and body size and the biotic factor's distribution range

In general, we found that the contribution of the biotic factor to the modelled species *increased* as the geographic range of the biotic factor *decreased* and as flower specialization *increased*. From the models of the pollen-collecting bees (Appendix text 3 in the electronic supplementary material). the model with the lowest AICc (AICc = 960.8) resulted in the formula with flower specialization (β1) and distribution range of the biotic factor (β2): y = (0.063 – 0.0099 * β1 + 0.013 * β2)$^{-1}$ (R^2 = 0.39; fig. 6A & 6B). This model was less than 2 AICc units away from the second best model that also included the body size, but the body size coefficient was not significant in this model (Appendix text 3 in the electronic supplementary material). The selected model of the cleptoparasitic bees contained the distribution range of the host bee (β) and it had the formula: y = $(0.13 + 0.0099*\beta)^{-1}$ (R^2: 0.47) (fig. 6C; Appendix text 3 in the electronic supplementary material). This model was less than two AICc units away from models that also included both or either body size and host specialization, but the coefficients of these variables were not significant (Appendix text 3 in the electronic supplementary material). When the response variable was changed to permutation importance, the trend between explanatory and response variables was similar and the same explanatory variables were selected, when using the AICc (Appendix figure 2 in the electronic supplementary material). This was the case for both pollen-collecting bees and cleptoparasitic bees.

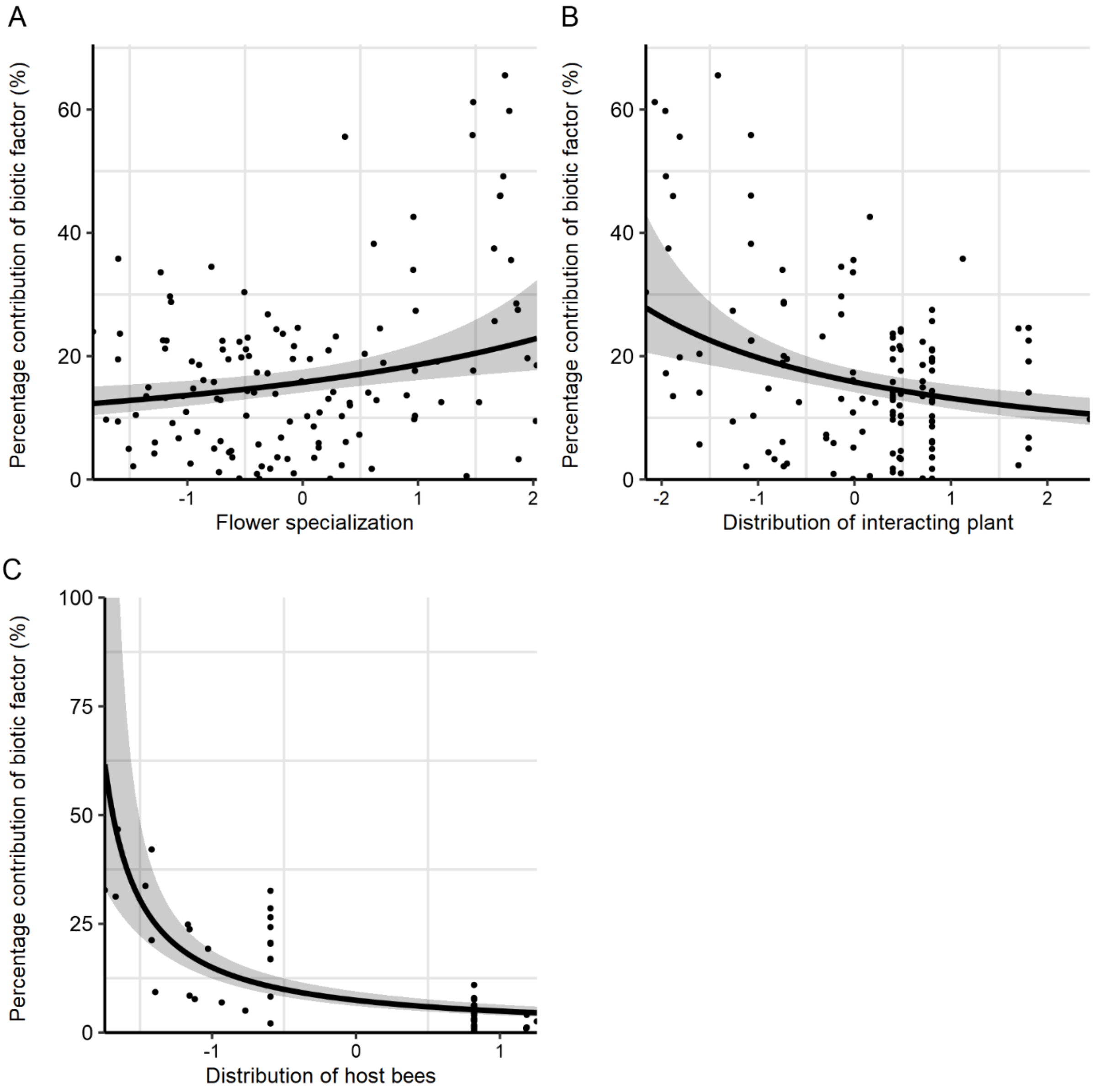

Figure 6: The results of the Generalized Linear Models (GLMs) show the relation between flower specialization (the inverse of the Shannon-Wiener index of number of plants genera interacted with) and the percentage contribution of the biotic factor to the models of the oligolectic and polylectic bees (Shannon 1948; fig. 6A). Fig. 6B shows effect of distribution of the most visited genus on the

contribution of the biotic factor to the model. Fig. 6C shows the relation between the distribution of the host bees and the contribution of the biotic factor to the models of the cleptoparasitic bees.

3.4 Discussion

Here we show that the inclusion of known interactions as biotic factors in SDMs improved our ability to explain the distributions of wild bees in the Netherlands. Adding the biotic factors to the abiotic models improved model performance for multiple evaluation measures. Additionally, the RI-SDMs showed that the improvement observed in the models was more common for the known interaction compared to any interaction.

Spatial resolution, taxonomic rank and distribution of the biotic factor all influenced the importance of the biotic factor to the models. A higher degree of specialization was correlated with a higher variable importance of the biotic interaction to the models. The model performance peaked at finer resolutions of the biotic factor up until 5 km for the cleptoparasitic bees and around 1 km for the pollen-collecting bees. The biotic factor at genus taxonomic rank was generally contributing more at finer resolutions and at species level than at coarser resolutions. A smaller distribution range of the biotic factor was also correlated with a higher importance of the biotic factor to the models.

Including biotic information in SDMs generally improves model performance

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 then polylectic bees. An increase in model performance by the addition of the host of the cleptoparasitic bee has been observed (Giannini 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), highlighting the importance of resolution in SDMs, particularly when including biotic information (Wisz et al., 2013). 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-Bégné et al. 2021) and biotic interaction can play a role in the distribution range edges of species even at a larger scale (Paquette and Hargreaves 2021, Freeman et al. 2022). Anderson (2017) suggested that biotic interactions can be unlinked variables that are independent from the modelled species. We argue that in our study the effects of the plant on the individual pollinator are negligible. Specialist and generalist pollinators often pollinate the same plant species and therefore represent a redundancy in the network (Scheper et al. 2014b). Additionally, an important cause of bee decline is the decline of their pollen host plants, while plants seem more threatened by abiotic variables (Scheper et al. 2014b), which suggests that the plants would be less affected by the individual pollinator. In the case of the cleptoparasitic bees, the host bee is affected negatively and we argue that by using a large time-scale and different resolutions these processes at a smaller time and geographical scale have a minimal effect on the models. A small percentage (2.42%) of the biotic factors were correlated with the soil variables that were the lowest contributing variables, and the contribution of these biotic factors to the models may be partly shared with soil requirements.

The RI-SDMs with interactions of random pollinated plants or host bees, revealed how the specificity of the interaction (e.g. specialist versus generalist) influences how a biotic interactions could be included in a SDM approach. 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 biotic factor in the models of the cleptoparasitic bees confirmed how important their host species are for modelling their distribution. Another contributing factor may be the biases in the data sources: the distribution of the cleptoparasitic bees and their hosts are sourced from the same wild bee occurrences database and therefore, share similar collection biases. In contrast to the distribution of the plant species which likely have their own separate collection biases. The RI-SDMs enable us to compare different randomized interactions that share similar collection biases. Our findings show that models of known interactions improve significantly more than randomized interactions for both cleptoparasitic bees and oligolectic bees. Therefore, it is unlikely that the similar collection bias of the modelled species and their known interaction is the primary factor influencing model improvement. In the case that data on the biotic factor 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, especially for generalist species (polylectic bees as opposed to oligolectic bees). A model improvement of randomized interactions could indicate that for this species an optimal foraging area can include a wider array of flowering plants. Another possibility is that the presence of the plant species could indicate other favourable abiotic conditions that are not explained by the abiotic factors. When multiple plant species improve model performance, the biotic factor could be included as an approximate of those plant-pollinator interactions by integrating the plant species as flowering plant diversity or diversity of the top five of most contributing plants to the model of the bee.

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 and 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 movement 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). The optimal spatial resolutions are similar to the recorded movement ranges of many wild bees, showing an optimal resolution of the biotic factor at finer resolutions from 100 m to 5 km for the cleptoparasitic bees and around 1 km for the oligolectic and polylectic bees.

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

The biotic factor had a higher contribution when created using genus level observations of the known interaction, at finer resolutions (500 m and 1 km). Using observation data at the genus taxonomic level could be a compromise, balancing the reduction in taxonomic resolution of biotic interactions with an increase in the number of records at finer resolutions. Additionally, pollinators generally visit closely related plants more often than would be expected by chance (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 conservationism, 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 2021), resulting in a dependency on multiple plant species in the same genus or family. 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 was found to be a good approximation of the biotic interaction. The cleptoparasitic bees showed lower variable importance of the biotic factor at coarse resolutions at the genus level compared to the biotic factor 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 biotic factor 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 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 are thought to parasitize on only one species (Bogusch et al. 2006).

The sampling effort for large observation databases is uneven across countries and continents (Beck et al. 2014). This is also true for wild bees, however the Netherlands is comparatively a well-sampled country (Marshall et al. 2024). This study suggests that when data is lacking, replacing biotic interactions with genus-level taxonomic data or using the species richness of multiple species as a proxy are good alternatives. Additionally, large data infrastructures like the Distributed System of Scientific Collections (DiSSCo; Hardisty et al. 2021) can help establish whether the sparsity of records is due to under sampling or a representation of the species' distribution. In some cases, data may not be missing but rather reflect the limited distribution and abundance of the biotic interactions. Whether data is missing or the biotic interaction has a very limited distribution and abundance, these interactions can still be incorporated into models. An absence of model improvement is not proof of an absence of an ecological interaction; additionally our study found a correlation between a higher importance of biotic interactions and a more limited extent of occurrence of the biotic interaction.

Influence of flower and host specialization and the distribution range of the biotic factor

We found that a higher degree of flower specialization for the flower visiting bees and a narrower distribution range of the biotic factor 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 at macroecological scales (Araújo and Luoto 2007). If the distribution of the biotic factor is narrow, it is more likely to be a limiting factor, delimiting the boundaries of the potential distribution of the focal species. Other studies have found that narrow distributions, although not too narrow, lead to more accurate models and high importance for certain key habitat factors (Tsoar et al. 2007, Syphard and Franklin 2010) and that specialist species yield better models than generalist species (Grenouillet et al. 2011, Marshall et al. 2015). We found that if the biotic relationship is strong, e.g. parasite-host relationships, then that becomes by far the most important factor.

An important consideration is that the biotic factor may be 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 are Joint-SDMs (Kissling et al. 2012). They are 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). However, the risk is high 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 et al. 2014, Ovaskainen et al. 2017). In this study the risk that any detected relationship between species is related to abiotic factors is lower for reasons that we integrate the biotic factor as the interaction known *a priori*, we compare the known interaction to randomized interactions, we tested for collinearity and we included a wide range of abiotic variables. Still, SDMs and Joint-SDMs estimate co-occurrence and correlations (Pollock et al. 2014) and to confirm the causality (true interaction) of the co-occurrence between species we would need to employ field studies. Here, we showed that biotic factors can improve the SDMs of wild bees in the Netherlands, especially when the distribution of the biotic factor is narrow and the modelled species is a specialist. Resolution and taxonomic rank of the biotic factor, should be taken into account to achieve the most optimal models. Our hypothesis that a biotic factor with a more narrow range would lead to a higher importance of this biotic factor to the models of the specialist species was confirmed. We recommend using single species or genus data as a biotic factor in the models of specialist species and to use an approximate, such as flower richness, for more generalist species.