UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole de biologie

ASSESSING THE RELIABILITY OF CITIZEN-SCIENCE DATA FOR THE STUDY OF ANT SPECIES' ENVIRONMENTAL NICHES AND DISTRIBUTIONS

Travail de Maîtrise universitaire ès Sciences en comportement, évolution et conservation, spécialisation « Ecologie et Evolution Computationelles»
Master Thesis of Science in Behaviour, Evolution and Conservation, specialisation « Computational Ecology and Evolution»

par

Marianna TZIVANOPOULOU

Directeur : Prof. Antoine Guisan
Superviseur (s) : Dr. Olivier Brönnimann
Expert (s) : Anonymous
Département d'écologie et évolution

Février 2023

ABSTRACT

To obtain data at large spatial and temporal scales, scientists are increasingly taking advantage of citizen-science projects, where data collection is done by volunteers. However, the potential presence of bias due to differences in species detectability and sampling effort in space could affect the interpretation of scientific results. Quantifying such bias for different taxonomic groups is therefore key to using the full potential of citizen-science datasets. Here, we compare the environmental niches and predicted distributions of ant species distribution models calibrated using scientific data or opportunistic citizen-science data. We find significant overlap between the quantified environmental niches and high correlation between model predictions for the majority of species. Divergence in model predictions was observed mostly for species with low detectability or which occurred in both natural and urban habitats. Based on these findings we developed a method to correct for the spatial sampling bias of the citizen-science dataset. Applying this bias correction increased the niche overlap and prediction correlations. Our findings indicate that citizen-science data can reliably be used for species distribution modelling, as long as the characteristics of the species studied are considered. Integrating scientific and corrected citizen-science data is recommended to minimize the risk of biased predictions.

Keywords: Citizen-science, Spatial bias, Niche overlap, Formicidae, Species distribution models

RESUME

Pour obtenir des données à larges échelles spatiales et temporelles, les chercheurs profitent de plus en plus des sciences participatives, qui permettent d’impliquer des volontaires dans la collection de données. Cependant, la présence potentielle de biais à cause de différences dans la détectabilité des espèces et l’effort d’échantillonnage dans l’espace pourrait influencer l’interprétation des résultats scientifiques. La quantification de ces biais pour différents groupes taxonomiques est donc essentielle afin d’utiliser pleinement le potentiel des données de sciences participatives. Dans ce rapport nous comparons les niches environnementales et les distributions de fourmis prédites par des modèles de distributions d’espèces calibrés avec des données scientifiques ou de sciences participatives. Nous observons qu’il existe un chevauchement significatif entre les niches environnementales quantifiées, et les prédictions des modèles étaient hautement corrélées pour la majorité des espèces. Des divergences entre les prédictions des modèles ont été observées principalement pour des espèces avec une détectabilité réduite ou vivant à la fois dans des habitats naturels et urbains. En se basant sur ces résultats, nous avons développé une méthode pour corriger le biais spatial d’échantillonnage dans le jeu de données de sciences participatives. Cette correction du biais a augmenté le chevauchement des niches écologiques et la corrélation entre les prédictions des modèles. Nos résultats indiquent que les données des sciences participatives peuvent être utilisées de façon fiable pour créer des modèles de distribution d’espèces tant que les caractéristiques des espèces étudiées sont considérées. Une intégration des données scientifiques et de sciences participatives est recommandée pour minimiser le risque de prédictions biaisées.

INTRODUCTION

Current research in biodiversity and conservation often requires ecological data at such spatial and temporal scales that cannot feasibly be collected by professional scientists alone. To fill these gaps, researchers are increasingly taking advantage of scientific projects which involve citizens collecting large quantities of data in a cost-effective way (Amano et al., 2016; Hochachka et al., 2012). The last decades have seen a huge growth in environmental citizen-science projects globally; in 2015 more than 1.3 million volunteers contributed in-kind as much as 2.5 billion dollars (Theobald et al., 2015). Data collected by citizens have already been used to detect the spread of invasive species, monitor biodiversity at international scales and investigate the impact of climate change on species phenology and range shifts (Chandler et al., 2017; Cooper et al., 2014; Dickinson et al., 2010; Larson et al., 2020). Citizen-science has the potential to increase the information available for under-studied areas and taxonomic groups, such as invertebrates (Callaghan et al., 2021). Apart from its contribution in increasing ecological knowledge, citizen-science can also promote public engagement in conservation and inform management planning (McKinley et al., 2017; Tulloch, Possingham, et al., 2013).

Citizen-science data are also increasingly being used for Species Distribution Modelling to study species distributions at local or global scales (Feldman et al., 2021). Species Distribution Models (SDMs) relate information on species occurrences to environmental predictors to predict habitat suitability (Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000). They can be used to address a large variety of ecological problems, including systematic conservation planning, modelling the impact of land-use change or predicting the expansion of invasive species (Elith & Leathwick, 2009). Species data used in SDMs can involve information about the presence-absence of a species, presence-only or abundance. Presence-absence data follow a more robust sampling design and provide information on sampling effort and bias, but their availability is often limited (Phillips et al., 2009). On the other hand, resources such as the Global Biodiversity Information Facility (GBIF) have made available millions of species presence records, with citizen-science projects such as eBird and iNaturalist being among the most important contributors (Gaiji et al., 2013; GBIF: The Global Biodiversity Information Facility, 2022). It is therefore not surprising that the rate of growth in the number of publications using citizen-science data for species distribution modelling is two times larger than the studies using SDMs in general (Feldman et al., 2021). Additionally, there has been great interest recently in developing models that integrate citizen-science and scientific datasets while considering the advantages and disadvantages of each approach (Fletcher Jr. et al., 2019; Isaac et al., 2020; Miller et al., 2019).

One weakness of citizen-science data that needs to be accounted for is the frequent presence of different forms of bias and errors that lead to lower use by traditional science in publications (Anderson et al., 2020; Gardiner et al., 2012; Graham et al., 2004; Theobald et al., 2015). Compared to professional scientists, volunteers may be able to detect fewer species and underestimate the abundance of certain taxa (Kremen et al., 2011). Identification errors may also be more common in citizen-science data for specific taxonomic groups such as insects (Hochmair et al., 2020). Many popular citizen-science projects that rely on massive participation do not have well-defined sampling protocols, promoting the opportunistic collection of data by volunteers (Pocock et al., 2017). These datasets are often characterized by the presence of temporal or spatial bias in sampling. For example, volunteers may visit more frequently sites close to highly populated areas or regions known for their biodiversity and the presence of rare species (Botts et al., 2011; Mair & Ruete, 2016; Tulloch, Mustin, et al., 2013). These biases are a challenge to the future potential use of citizen-science data in SDMs as their impact on ecological inferences remains still unclear (Johnston et al., 2022).

To better understand the sources of bias and assess the reliability of citizen-science data for species distribution modelling, scientists have been comparing the performance and predictions of SDMs calibrated with citizen-science or scientific datasets, such as survey, satellite telemetry or camera trap data, with contrasting results. Although some studies have shown that model performance and predictions are comparable between scientific and citizen-science data (Coxen et al., 2017; Tanner et al., 2020; Tye et al., 2017), in other cases models calibrated using citizen-science datasets tended to overestimate the habitat suitability of highly populated areas and neglect sites of interest for a given species (Planillo et al., 2021). Tiago *et al.* (2017) compared ecological niches and species distributions of reptiles and amphibians modelled using citizen-science and scientific data. They found that the simulated niche overlap and model performance presented great variability depending on the species. These results highlight the need to investigate more broadly the impact of citizen-science sampling biases in other taxonomic groups, especially as previous work has only focused on vertebrates. They also emphasize the importance of improving methods to cope with bias in citizen-science data.

In this study, we assess whether citizen-science data can reliably be used for ant species environmental niche modelling. For this purpose, we compare the environmental niches and predicted distributions of 15 selected ant species calibrated using either opportunistic citizen-science data or structured scientific data. Moreover, we propose a method to correct for the environmental bias present in citizen-science data and apply it to generate maps of ant habitat suitability in the canton of Vaud, Switzerland, by pooling the scientific and the corrected citizen-science dataset. Our results confirm the presence of species-specific differences with regards to the impact of sampling bias in citizen-science data and indicate the potential of the proposed method to mitigate them.

MATERIALS AND METHODS

Study Area

Ant species data were collected in the canton of Vaud, located in western Switzerland (46.2–47.0^{o}N, 72 6.1–7.2^{o}E) and covering a surface of 3,211.94 km^{2}. The area is topographically heterogeneous with an altitude ranging from 372 m to 3,210 m. The hilly central plateau is flanked in the west and the east by the mountainous regions of Jura and the Alps, respectively. Located in the south of the canton, on the shore of Lake Geneva, Lausanne is the largest city with a population of 140,000.

Species Data

Citizen-science data of ant diversity in the canton of Vaud were retrieved from the "Opération Fourmis" citizen-science project (https://wp.unil.ch/fourmisvaud/), which took place between spring and autumn 2019. Sampling kits were distributed to interested volunteers and ants were collected opportunistically without a structured sampling protocol. Volunteers were encouraged to collect 10 ant workers from the same colony, preferably from a nest and place them in vials filled with ethanol. The vials with the sampled ants, along with information on the date and location of the sampling, were sent to the University of Lausanne, where ants were identified at the species level morphologically and genetically by experts (see Avril et al., 2019).

During the same period, ant diversity inventories were carried out by researchers from the University of Lausanne following a random stratified approach (Table 1, Freitag et al., 2020). Sampling took place in 44 pre-established monitoring sites of 1 km^2 across the canton of Vaud. In each site, 25 plots were randomly chosen proportionally to the surface covered by every habitat type within the site, with at least one plot for each habitat. In case part of the site was not accessible (for example if covered by water), the number of selected plots was reduced proportionally to the inaccessible surface. For each plot, ants were sampled at 6 points located evenly around the circumference of a 2 m circle with the center of the plot as its center. The presence of ant nests under rocks, downed wood and 2 L of litter, soil and small rocks was investigated at each of these 6 points. Moreover, ant workers were also collected from all trees in the circle with a diameter larger than 3 cm at breast height. Finally, in each sampling site 2 km of transects were surveyed and ants were sampled from mounds within 2 m of the transect line. Transects were mapped in a way so that all habitats within the sampling site were proportionally represented. By combining these 3 sampling strategies the collection of ants with different nest construction methods was possible. Soil samples and tree samples targeted ant species excavating nests underground or on trees respectively, while transects were more adapted to species creating aboveground nests such as the colonies of *Formica* wood ants. From each ant colony, 10 workers were placed in ethanol vials and ants were identified to the species level as described above (Szewczyk et al., 2022, Preprint).

Table 1. Habitat types used to define the random stratified sampling strategy. The percentage of the total area of the canton of Vaud, and of the total area surveyed during sampling, covered by each habitat is given.

Habitat type	**Proportion of total surface in the canton of Vaud (%)**	**Proportion of total surface in the sampled 1 km^2 sites (%)**
Coniferous forest	13.39	11.31
Deciduous forest	8.07	9.44
Urban areas	7.84	5.30
Mixed forest	7.10	5.88
Edge	6.29	7.08
Transportation	4.66	3.96

Permanent agriculture	2.14	1.87
Scree	1.00	2.04
Dry meadow	0.87	3.47
Alluvial zone	0.55	0.55
Swamp	0.52	0.67
Humid zone	0.09	0.14
Gravel pit	0.08	0.13
Quarry	0.02	0.06
Clay pit	0.01	0
Other open habitats*	47.37	48.10

** Including pastures, meadows, lawns, and crops (not in the permanent agriculture category)*

Overall, 6,632 samples from 76 species were collected during the citizen-science campaign by 606 collectors (Freitag et al., 2020). After removing samples with low accuracy of geographic coordinates, 6,384 observations were retained. The scientific inventories yielded a total of 1,453 observations from 51 species, of which 1,090 were found in the soil, 343 along transects and 20 on trees. In order to have sufficient data for the implementation of the species distribution models, only species with 20 or more occurrences in each of the two datasets were kept for the remainder of the study. Below this sample size, model performance declines and models become prone to over-fitting with reduced reliability when predicting in new areas or environmental conditions. (Guisan et al., 2017). This resulted in a list of 15 species: *Formica cunicularia, F. fusca, F. lemani, F. lugubris/paralugubris, F. pressilabris, Lasius flavus, L. niger, Myrmica rubra, M. ruginodis, M. sabuleti, M. scabrinodis, Solenopsis fugax, Temnothorax nylanderi, Tetramorium caespitum* and *T. impurum*. It should be noted that *Formica lugubris* and *Formica paralugubris* are two distinct species, but because it was not possible to distinguish them morphologically during identification, and their habitat preferences are very similar, they were considered together in this and previous studies (Szewczyk et al., 2022, Preprint). As the number of occurrences for some of the chosen species was low in the scientific dataset (24 for *F. cunicularia*, 25 for *M. rubra*), additional sampling was carried out in spring-summer 2022 in the context of the current study at 4 sampling sites following the protocol described above. The additional sampling sites were chosen to focus on habitats that were proportionally under-sampled in the original scientific inventories, such as urban areas, permanent agriculture, and coniferous forests. 131 samples were collected and identified morphologically for the *Formica*, *Lasius, Solenopsis* species, as well as *M. rubra* and *M. ruginodis* and genetically for *Temnothorax* and *Tetramorium* species, *M. sabuleti and M. scabrinodis* (see Supplementary Materials and Methods).

Environmental data

Environmental predictors for modelling were chosen based on expert knowledge of environmental factors influencing ant distribution and diversity (Seifert, 2018), as well as previous research on ant species distribution models (Chen, 2008; Fitzpatrick et al., 2013; Hartley et al., 2006; Ward, 2007). Climatic variables at 25 m x 25 m resolution were extracted from the CHClim25 dataset, including precipitation, temperature, growing season and Worldclim bioclimatic variables (Broennimann, 2018). The aspect and slope were calculated from a Digital Elevation Model using the *terrain* function from the raster package in R (Hijmans et al., 2022). Land-use variables included the proportion of each pixel covered by crops, forests, edges, or urban habitat. To take into account the mobility of ants (Seifert, 2018), the proportion of each land-use category was also calculated in a 25 m and 200 m neighborhood around each pixel using the function *focal* from the raster package. Anthropogenic pressures were represented by the length of roads and the perimeter of buildings in each pixel using information derived from OpenStreetMaps. Information about soil conditions was captured using plant Ecological Indicator Values in each pixel (Descombes et al., 2020). As a measure of vegetation productivity, the average canopy height by pixel and the mean NDVI during the period 2010-2020 were used. The average NDVI over the study area was calculated from 206 images captured by LANDSAT 5 and LANDSAT 8 using Google Earth Engine. Overall, an initial set of 59 predictors with a resolution of 25 m x 25 m over the study area was assembled. Variable pre-processing was performed in R 4.1.1 (R Core Team, 2021) and QGIS 3.16.13.

Ecological Niche Quantifications

To quantify the ecological niche of ant species using each dataset separately, the approach of Broennimann et al., 2012 was followed using the ecospat package v3.4 (Broennimann et al., 2022; Di Cola et al., 2017). In short, an environmental Principal Component Analysis (PCA) was calibrated over all the pixels of the study area using the 59 predictor variables, and the first two principal components were used to define the environmental space over the canton of Vaud, which was divided into a 100 x 100 cell grid. The density of species occurrences in each grid cell was calculated by applying a kernel density function to the number of occurrences across the environmental space separately for the two datasets, without excluding marginal sites with regards to species observations or environmental conditions (threshold = 0). The overlap between the species niches as inferred by each dataset was measured using Schoener's D statistic, where 0 indicates no overlap and 1 complete overlap. To assess whether the ecological niche modelled with the citizen-science dataset is more similar to the niche modelled with the scientific dataset than expected by chance, the similarity tests were performed with 1000 repetitions. For each repetition, the observed density of species occurrences in the environmental space was randomly shifted for one of the two datasets, and the overlap between this simulated niche, and the observed niche modeled with the other dataset was calculated. The result of the test is significant if the true overlap of the niches modeled with the two datasets is greater than 95% of the simulated overlap values. The overall size of the ecological niche of each species was measured based on the number of grid cells in the environmental space occupied by the niche of each species. The proportion of the ecological niche covered by each dataset was quantified by calculating the niche stability (i.e., proportion of the environmental niche covered by both datasets), unfilling (i.e., proportion of the niche covered by the scientific dataset only) and expansion (i.e., proportion covered by the citizen-science dataset only) metrics (see Guisan et al., 2014) with the ecospat package.

Species Distribution Models

As a trade-off between the need to accurately represent the ecological niche of each species on the one hand, and the limits imposed by the computation time and the low number of occurrences for some species on the other hand (90 observations by species on average in the scientific dataset, see Table S1, Supplementary Materials and Methods), it was decided that for each species 8 predictors would be chosen to fit the Species Distribution Models (SDM). The nsdm package (Adde et al., In review) was used to carry out predictor selection after comparison against two other alternative approaches (Elastic Net Generalized Linear Models (GLM) and PCA, Table S3, Supplementary Materials and Methods). The nsdm method is based on an ensemble approach, which consists of excluding colinear variables with a correlation higher than 0.7 and then combining Generalized Additive Models, Random Forest, and Elastic Net GLM algorithms to rank the remaining variables based on their importance. Following Barbet-Massin et al. (2012), 10,000 background points, which were randomly selected in the canton of Vaud, were added to the occurrences of both datasets for each species and the two classes, presences and background points, were weighted equally. The 8 predictors with the highest ranking for each species were retained (Table S4, Supplementary Materials and Methods).

The distribution of each ant species in the canton of Vaud was modelled using the citizen-science and the scientific dataset independently along with the set of 10,000 randomly selected background points that was also used for predictor selection. Because of the small amount of data for some of the ant species selected for modelling, the Ensemble of Small Models (ESM) approach from the ecospat R package was used to predict the distribution of ant species over the canton of Vaud, as it has been shown to perform better than conventional SDMs for rare species (Breiner et al., 2015; Lomba et al., 2010). This method consists in calibrating multiple small models with two predictor variables and averaging their predictions based on the performance of each model during evaluation. To fit the bivariate models, the Generalized Boosted Models (GBM) algorithm was used, which is recommended to produce ESMs with high predictive performance (Breiner et al., 2018). The parameters of the GBM models were fixed according to the default biomod2 settings of 2500 trees, a shrinkage of 0.001, a bag fraction of 0.5 and 3 cross-validation folds, which is within the recommendations of Guisan et al. (2017). The interaction depth was set to 2, following the instructions of the ecospat package. To evaluate the performance of the models, 19 split-sample evaluation runs were carried out using 70% of the data for model training. Models were evaluated based on the Area Under Curve (AUC), the maximum value of the True Skill Statistic (MaxTSS), and the Boyce index. The AUC was used to weight the bivariate models and create the ensemble predictions. Models with an AUC lower than 0.5 were completely excluded from the ensemble predictions as SDMs with a performance bellow this threshold were counter-predicting species' habitat suitability.

Post-hoc Statistical Analyses

To compare the predictions of the SDMs using the two datasets, the difference in the occurrence probability of each pixel as predicted by the models created either with the citizen-science dataset or the scientific dataset was measured. Moreover, the correlation between the habitat suitability predicted by each dataset at each pixel over the study area was also calculated using Spearman’s correlation index. The proportional contribution of each predictor in the final ensemble ESM species distribution models for each species was obtained using the function *ecospat.ESM.VarContrib* from the ecospat package. For each variable, the ratio of the sum of AUC of each bivariate model containing the predictor against the sum of AUC of all bivariate models without the predictor was calculated and the result was corrected based on the number of bivariate models of each ESM with or without this predictor. A ratio higher than 1 indicates that a given predictor improves the performance of the model when added compared to the average.

To investigate potential interspecific differences in the performance of the SDMs and the correlation between the predictions of the models using the scientific or the citizen-science dataset, the relationship between these variables and the species’ niche size was assessed using the lme4 package in R (Bates et al., 2015). After controlling for the distribution of the response variable and the residuals, a gaussian distribution was assumed. Therefore, linear mixed-effects models were used with the niche breadth as the predictor variable, and as response variables for each species Spearman’s rho between the two dataset predictions, the AUC of the ensemble model for the scientific dataset, or the AUC of the ensemble model for the citizen-science dataset. This approach was chosen following Tiago et al. (2017) to account for the potential confounding effect of phylogenetic non-independence between the species, by adding the ant subfamily (Formicinae or Myrmicinae) as a random effect.

Correction of citizen-science data environmental bias

To characterize the sampling bias of the citizen-science dataset with regards to the environmental conditions in the study area, the environmental bias correction approach of Chauvier et al. (2021) was followed. First, a PCA was carried out using the 59 environmental predictors as described above, and the first 17 principal components, which explain 95% of the variance, were retained. A clustering analysis was carried out using the *clara* algorithm from the cluster package in R (Maechler et al., 2022) with the 17 principal components as predictors to separate the study area into 8 environmental clusters. The number of environmental clusters was chosen based on the optimal silhouette score (Rousseeuw, 1987) and the number of observations available for each species and dataset. The proportion of observations in the scientific dataset located in each environmental cluster was quantified and used to subsample the citizen-science dataset. The frequency of scientific observations in each environmental cluster was used to define the number of citizen-science observations to be sampled, so that, in the end, the proportion of observations by cluster would be the same between the scientific and the corrected citizen-science dataset. Eleven corrected subsets of the citizen-science dataset were created by randomly sampling the appropriate number of observations of the citizen-science dataset within each environmental cluster.

Previous modelling analyses were repeated using these corrected subsets of the citizen-science datasets to assess how the random choice of observations by cluster could influence the variability of model results. Environmental niche quantifications were performed using 10 corrected subsets and the niche overlap was quantified. Predictor selection with nsdm was done using only one corrected citizen-science subset along with the scientific dataset and the previously selected background points. The same set of variables was retained for all subsequent modelling to restrain additional model variability that could be due to randomness in the choice of predictors. Species distribution models were created for 10 different corrected citizen-science data subsets for each species except *F. cunicularia, M. scabrinodis and T. impurum,* as the number of observations in the sub-samples were not sufficient for modelling (see Supplementary Materials and Methods, Table S1). The last corrected subset was pooled with the scientific dataset to create integrated SDMs for all 15 species. The parameters of the SDMs were identical to those in the previous section, however for the models using only the corrected subsets, the number of cross-validation folds was reduced to 9 to limit computation time.

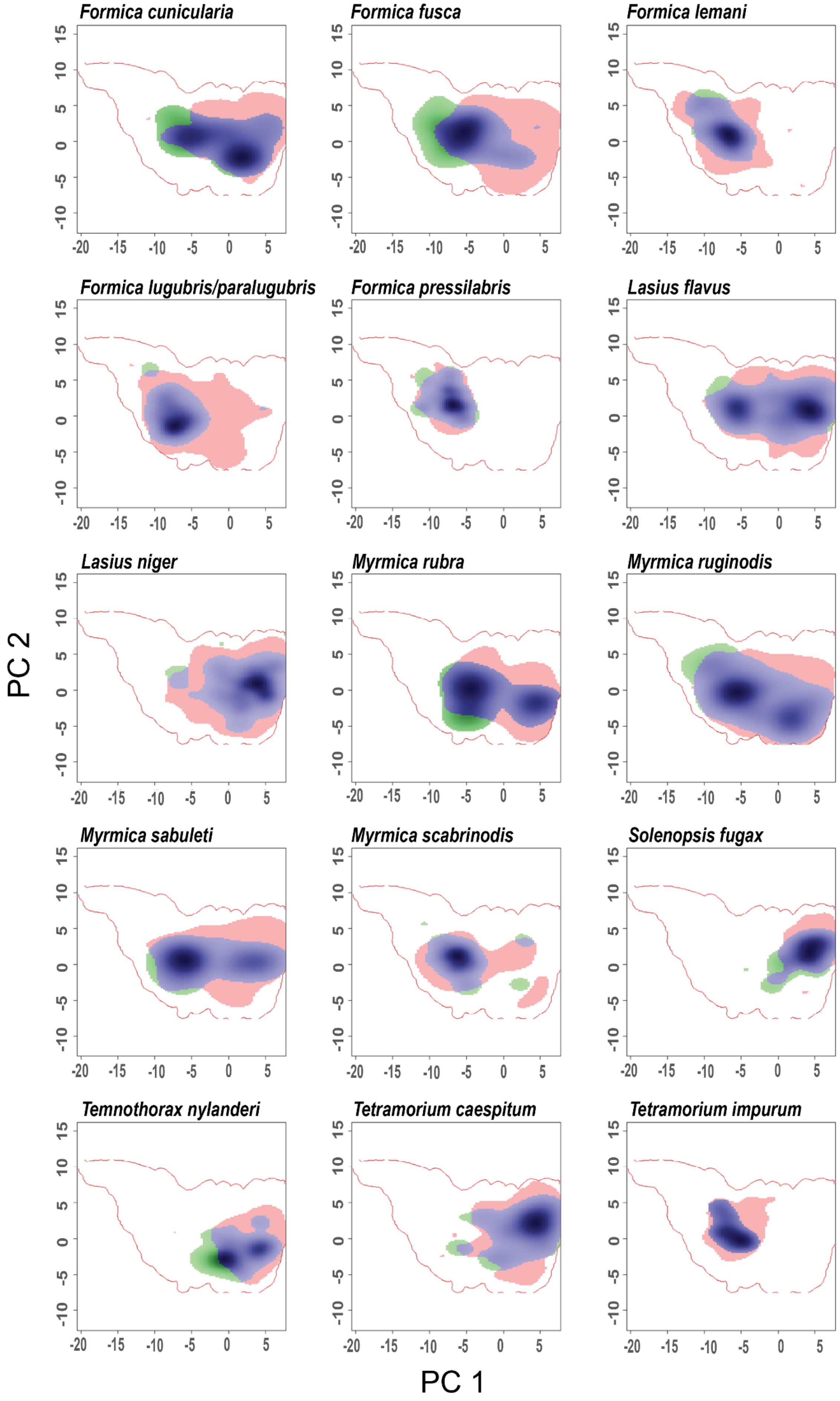
RESULTS

Ecological Niche Quantifications

Broadly speaking, we found high values of overlap between environmental niches modelled using either scientific or citizen-science data. The average overlap was of 0.64, with ecological niches simulated using the two datasets being significantly more similar than expected by chance for all 15 species (Supplementary Materials and Methods, Table S2). Seven ant species had a niche overlap equal or higher than 0.7 and only three, *Formica cunicularia*, *F. fusca* and *Temnothorax nylanderi* had values of overlap lower than 0.5 (0.344, 0.380 and 0.473 respectively).

The proportion of the environmental space covered by the niche simulated with the citizen-science dataset but not the scientific dataset was for all species higher than the proportion covered by the scientific dataset only (Figure 1). In fact, the percentage of the niche covered by the scientific dataset exclusively never surpassed 25% (mean: 0.065 ± 0.077), whereas as much as 48% of the niche of the species *F. fusca* was represented only by the citizen-science dataset (mean: 0.184 ± 0.116). After examining the coverage of the ecological niches in the environmental space, we observed that niches modelled with the citizen-science dataset tended to be located towards more positive values along the first PCA axis, associated with higher temperatures and lower precipitation (Supplementary Materials and Methods, Figure S1).

Figure 1 (next page). Results of ant species environmental niche quantifications. The axes represent the first two principal components of the environmental PCA. Red lines delimit the extent of the environmental space of the study area. The zones of the environmental space covered only by the ecological niche of each species simulated with the scientific or the citizen-science dataset are given in green and pink respectively. Areas covered by both datasets are in purple.


Formica cunicularia
Formica fusca
Formica lemani
Formica lugubris/paralugubris
Formica pressilabris
Lasius flavus
Lasius niger
Myrmica rubra
Myrmica ruginodis
Myrmica sabuleti
Myrmica scabrinodis
Solenopsis fugax
Temnothorax nylanderi
Tetramorium caespitum
Tetramorium impurum
15
10
5
0
-5
-10
-20
-15
-10
-5
0
5
PC 2
PC 1

Species Distribution Models

The performance of SDMs varied between species and depending on the dataset chosen for modelling, however, no model was poorly evaluated (mean AUC: 0.873 ± 0.061, Supplementary Materials and Methods, Table S5). Seven species had excellent model performance based on the AUC (values larger than 0.9) when using occurrences from the scientific dataset compared to 4 species when using the citizen-science dataset. Model performance was good (AUC between 0.8 and 0.9) for 4 species with the scientific and 10 species with the citizen-science dataset. Finally, the remaining 4 ant species using the scientific dataset and 1 species using the citizen-science dataset had a fair model performance (AUC between 0.7 and 0.8).

Ensemble model predictions of habitat suitability over the study area highlighted differences in environmental preferences between ant species in the canton of Vaud. For the sake of conciseness, the predictions of the models are presented for 4 selected species representing different challenges for data sampling by volunteers in Figure 2. *Formica pressilabris* and *Lasius niger* show more specific habitat preferences, the former occurring mainly in mountainous and alpine grasslands and the later in rural and urban areas (Seifert, 2018). *Formica fusca* occupies a wider range of land-use types including both natural and urban areas. Finally, *Temnothorax nylanderi* prefers broadleaved forests and can be difficult to detect due to its small size and nesting habits. The maps of habitat suitability predictions for the remaining species are given in the Supplementary Materials and Methods (Figures S2-S4).

The distributions predicted by models based on either the scientific or the citizen-science dataset were more consistent for species occurring preferentially in specific habitat types. For example, both predicted higher occurrence probabilities in the Jura mountains and western Alps for species *F. lemani, F. pressilabris* and *F. lugubris/paralugubris* that are known to occur in montane and alpine grasslands or forests. On the other hand, we found higher divergence between model predictions for more generalist species establishing in both natural and rural or urban areas. This can be observed for species *F. cunicularia* and *F. fusca* which occur in natural grasslands as well as rural or disturbed habitat in peri-urban areas. The scientific dataset predicted higher habitat suitability in hilly and mountainous areas while the citizen-science dataset predicted higher habitat suitability across the central plateau and in urban areas in the south.

Overall, maps of the difference between the predictions of the citizen-science and the scientific models revealed that the citizen-science dataset tended to over-predict the suitability of urban areas compared to the scientific dataset for many ant species, such as *L. niger* (Figure 2, right panels). We also detected this divergence in habitat suitability predictions when comparing the response curves produced by the models using the two datasets. Citizen-science models in which the percentage of urban habitat within or in the neighborhood of each pixel was chosen as a predictor indicated a rise in probability of occurrence when the percentage of urban area increased for the majority of species. On the other hand, models based on scientific data showed only moderate or no change in the probability of occurrence as a function of predictors related to urban habitat.

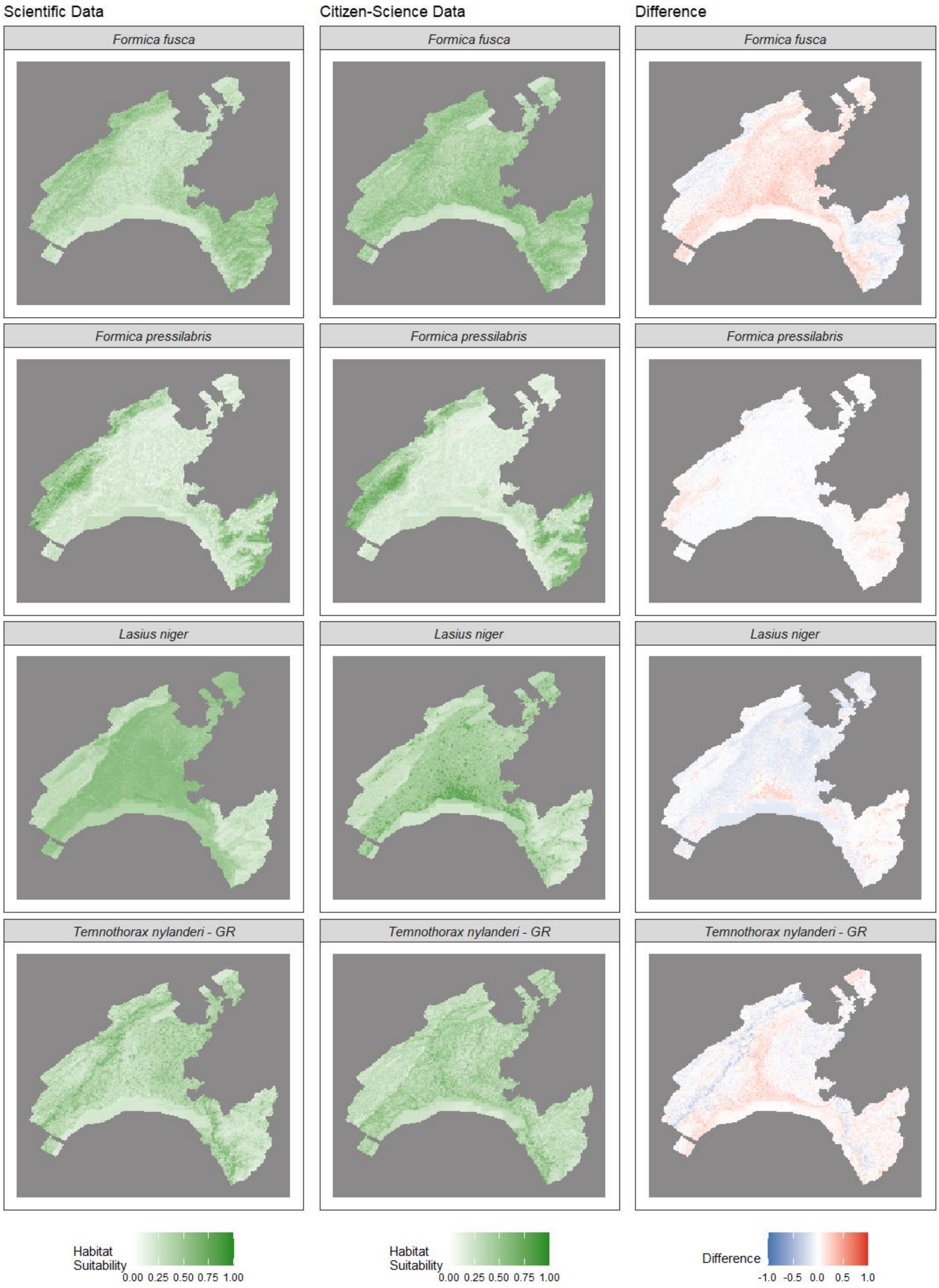


Figure 2. Results of ant species distribution models for 4 selected species: i) Predicted habitat suitability using the scientific dataset, ii) predicted habitat suitability using the citizen-science dataset, iii) difference between the predictions of the citizen-science models and the scientific models. Red color indicates that the predictions of the citizen-science models are higher than the scientific models.

Generally speaking, the response curves for the other environmental predictors chosen in this study showed comparable relationships between the models based on the scientific or the citizen-science dataset. The relative importance of model environmental predictors was also similar between datasets for the majority of the variables selected in the models (Figure 3). The predictors with the greatest relative importance for both datasets and across all species were the aspect, evapotranspiration and soil pH, however, the last two variables were only selected in the models for one species (*F. lemani* and *T. nylanderi* respectively). The importance of urban habitat as a predictor was higher for models using the scientific dataset compared to the citizen-science dataset, the performance of the majority of the bivariate citizen-science models including these predictors performed slightly worse than the average bivariate model within a given ESM.

Comparison of SDMs predictions between datasets and species

The correlations between the predictions of the SDMs using the scientific or the citizen-science data further confirmed the results obtained through the environmental niche quantifications. The habitat suitability in a given grid cell of the study area as predicted by the scientific dataset was positively correlated to the corresponding citizen-science prediction for all species with an average Spearman's rho of 0.74 (*p-value* < 0.05). Nevertheless, we identified large interspecific differences in the strength of the relationship (Figure 4). The correlation was higher than 0.75 for 11 out of the 15 ant species, with 5 species having a value higher than 0.9. Interestingly, species with the lowest niche overlap also showed the lowest values of correlation. Among those *F. fusca, M. sabuleti* and *T. nylanderi* had a correlation around 0.5. *F. cunicularia* had the lowest correlation value at 0.004. For these 4 species, cells where the scientific dataset predicted low habitat suitability tended to have low suitability in the citizen-science predictions as well, however areas with higher habitat suitability as predicted by the scientific dataset could correspond to a wide range of suitability values in the citizen-science predictions.

Interspecific differences in the correlation between dataset model predictions could not be explained based on the species' environmental niche breadth, but a strong relationship was detected with regards to model performance (Supplementary Materials and Methods, Figure S7). We found that Spearman's rho between model predictions of the scientific and citizen-science dataset was not significantly correlated to niche breadth quantified through environmental niche quantifications, although a slightly negative relationship was found (*adjusted* R^2 = -0.003488, *p-value* = 0.35). On the other hand, model performance measured by the AUC was negatively correlated to niche breath for both models created using the scientific or the citizen-science dataset (*adjusted* R^2 = 0.5579, *p-value* < 0.05 and *adjusted* R^2 = 0.8432, *p-value* < 0.05 respectively). SDMs performed therefore comparatively worse when modelling the distribution of species whose inferred environmental niche breadth was larger. Adding phylogeny as a random effect did not improve model fit. For all three models, variance between subfamilies was very low and simple linear models using only niche breadth as a predictor yielded a slightly better AIC.

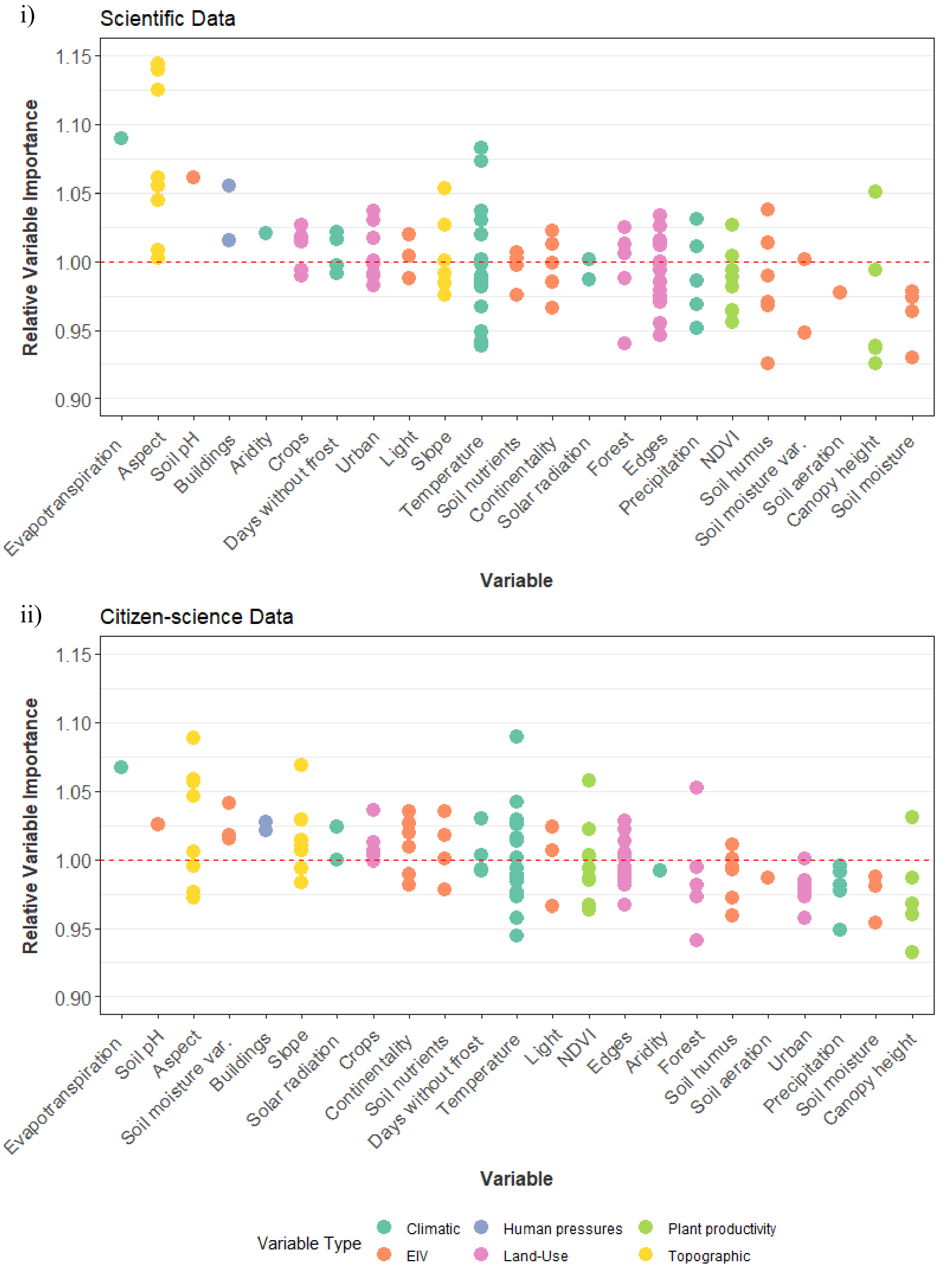


Figure 3. Relative importance of variables in the ant SDMs using i) the scientific dataset, ii) the citizen-science dataset. Each point represents the importance of the variable for one species. Variables related to temperature, precipitation and habitat types were combined to facilitate visualization.

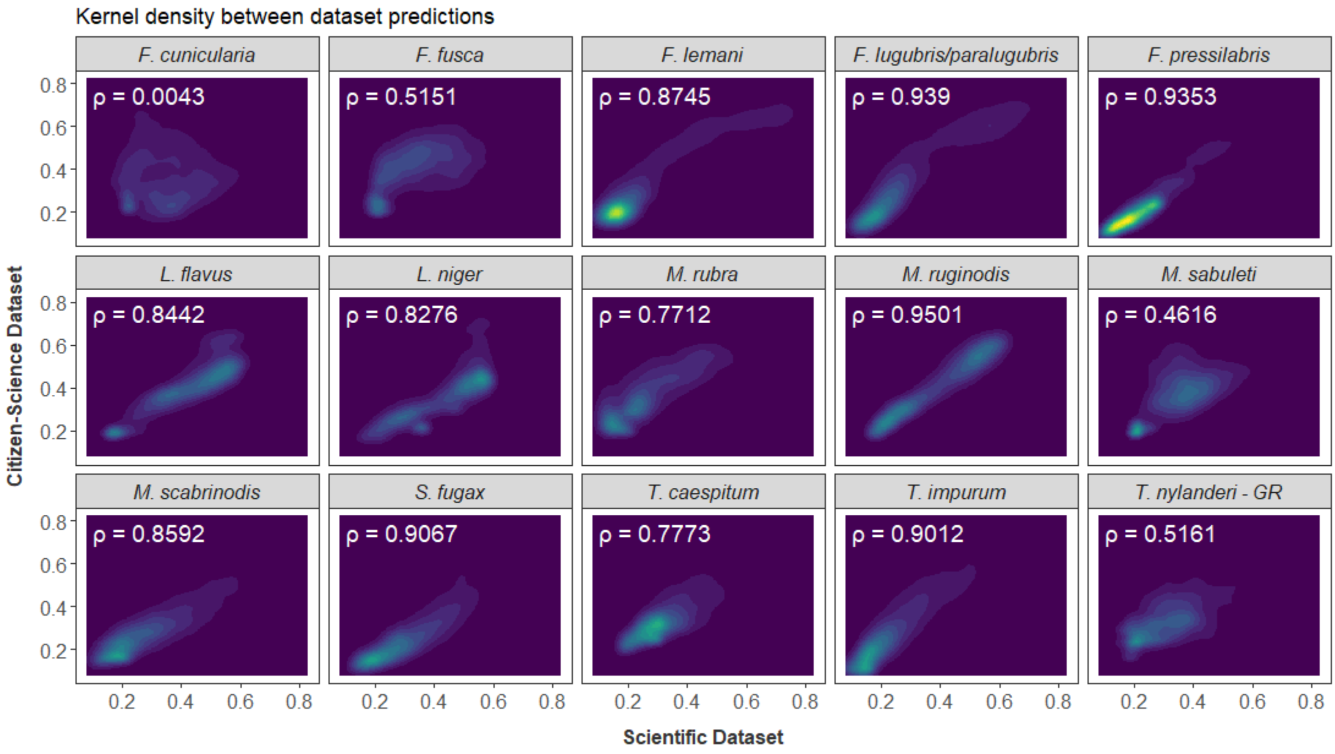


Figure 4. Relationship between the predictions of habitat suitability from the models created with the scientific (x axis) or the citizen-science dataset (y axis). Kernel density plots were built using the predictions at 10,000 random grid cells over the study area. Spearman's rho is given in white.

Correction of citizen-science dataset

Producing subsets of the citizen-science dataset corrected for environmental bias improved largely the overlap between the environmental niche quantified with the scientific dataset and the niche simulated with the corrected citizen-science dataset for all 15 species (mean: 0.740 ± 0.082, Figure 5). We found that the species with the lowest overlap when using the citizen-science dataset before correction were those that presented the highest increases in niche overlap after correction. As a result, 10 ant species had an average niche overlap larger than 0.7 and the remaining 5 species had an overlap higher than 0.5. Overall, the mean niche overlap increased by 0.1, from 0.64 to 0.74. Niche similarity between the scientific data and the corrected citizen-science subsets was statistically significant for the majority of species with the exception of *M. rubra*. Most quantified niches of this species using the corrected citizen-science subsets were not significantly similar to the niche based on the scientific dataset.

Species Distribution Models created using the corrected subsets of the citizen-science dataset were evaluated as "fair" or above (all AUC values > 0.7, mean: 0.895 ± 0.053). The consistency in the predictions of habitat suitability over the study area across corrected subsets varied between species (Supplementary Materials and Methods, Figure S5). Predictions were highly consistent for *F. lugubris/paralugubris*, *F. pressilabris* and *L. niger* as the standard deviation in the values of habitat suitability by grid cell in the canton of Vaud was overall low. For the other species, the highest values of standard deviation were located in areas with the highest predicted probability of occurrence. For example, areas of high elevation presented increased variation in habitat suitability for *F. lemani*, while for *S. fugax* the most variable areas were located along the central plateau and around urban habitats.

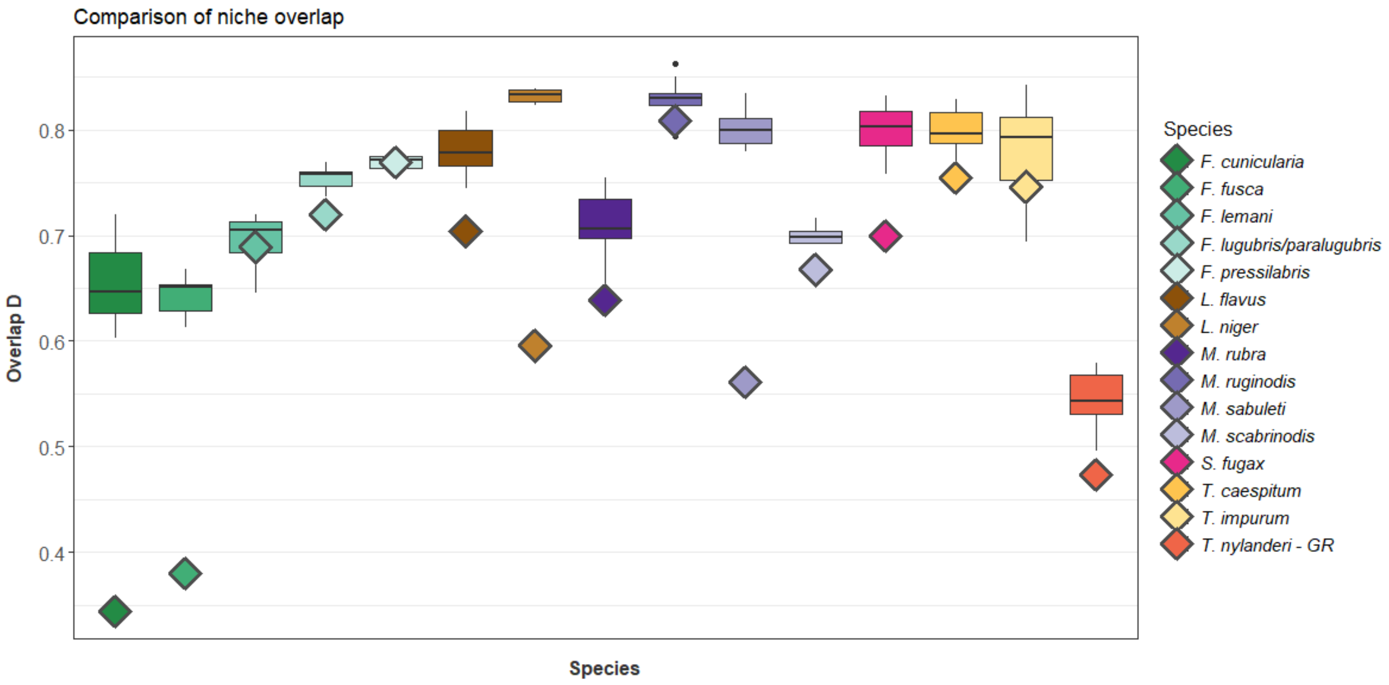


Figure 5. Overlap between ant environmental niches quantified using i) the scientific dataset and the citizen-science dataset (points), ii) the scientific dataset and 10 different randomly selected subsets of the citizen-science dataset corrected for environmental bias (boxplots).

Similarly to niche overlap, we also observed an improvement in the value of Spearman's correlation between scientific and citizen-science predictions for the majority of the species studied (Supplementary materials and Methods, Figure S8). Species for which the correlation between the predictions of the scientific and the uncorrected citizen-science dataset were low had the greatest increase (from 0.52 to 0.72 for *F. fusca* and 0.46 to 0.64 for *M. sabuleti*). However, we also recorded slight reductions in the values of the correlation between predictions for 4 of the 15 species. These included 3 species with a high correlation between scientific and uncorrected citizen science predictions (*F. pressilabris*, *M. ruginodis*, *S. fugax*), as well as *L. flavus* which witnessed the largest decline in Spearman's value of rho (from 0.84 to 0.79). *L. flavus* along with *T. nylanderi* also presented a larger variability in the value of correlation by corrected citizen-science subset compared to the other species.

Integrated Species Distribution Models

The predictions of habitat suitability across the canton of Vaud for the 15 ant species produced by pooling the scientific and the corrected citizen-science data are intermediate between those made by using each dataset individually (Figure 6, Supplementary Materials and Methods, Figures S2-S4). Although the predictions tended to be visually closer to the models based on the scientific dataset than on the uncorrected citizen-science dataset, they also highlighted areas that were identified by the citizen-science models as important for each species. For example, the probability of occurrence in peri-urban areas was comparatively larger in the predictions of the pooled dataset for species *F. fusca* and *S. fugax*. Similarly to models created using only one of two datasets, SDM performance was satisfactory for all 15 species with 5 species having an excellent performance, 8 species with a good AUC, and the remaining 2 species being evaluated as fair.

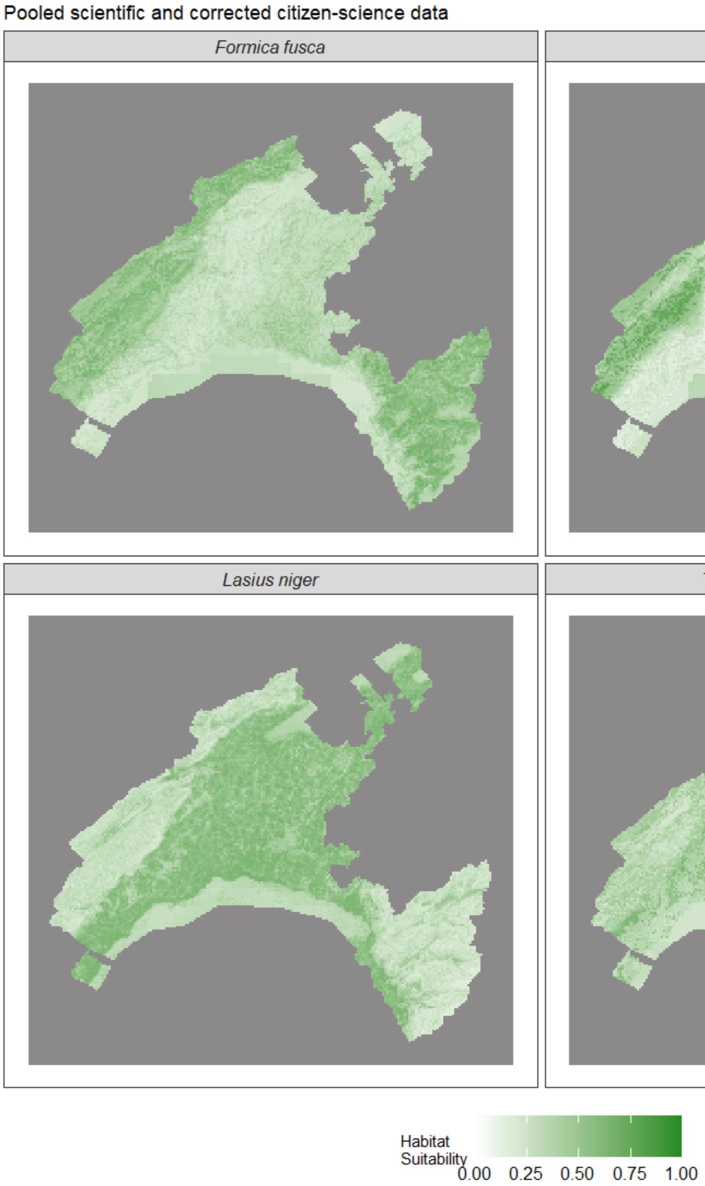


Figure 6. Results of ant species distribution models for 4 selected species using the scientific and the corrected citizen-science dataset

The relative importance of environmental predictors in the integrated models of ant species distributions was different from the models using either the scientific or the citizen-science dataset (Supplementary Materials and Methods, Figure S6). Although aspect remained the predictor with the highest relative importance on average, variables related to precipitation and soil properties such as nutrients and moisture variation played a larger role. The importance of land-use predictors declined when both the scientific and the corrected citizen-science data were considered. Variables related to the proportion of urban habitat were less often selected as predictors in the models and the response curves showed a more moderate relationship to the probability of occurrence compared to models using the citizen-science dataset before correction. Nevertheless, habitat suitability tended in general to increase along with the percentage of urban habitat for species *F. fusca, L. niger, S. fugax* and *T. caespitum*.

DISCUSSION

In this study, we assessed the reliability of using citizen-science data for species distribution modelling and showed that environmental niche quantifications and species distribution models based either on scientific or citizen-science data yielded comparable results for the majority of the 15 selected ant species. Despite the high values of niche overlap and correlation between model predictions, we found large species-specific differences in the consistency of analysis results between the two datasets. These differences are attenuated when applying the correction method, as there is a large improvement in both overlap and correlation values.

Comparison of scientific and citizen science model results

The high values of niche overlap and Spearman's correlation between the predictions of the models using either the scientific or the citizen-science dataset for the majority of ant species indicate that in general citizen-science data can be used to get reliable predictions on species distributions. Our results are therefore in line with previous findings for bird species, which demonstrated high overlap when comparing model predictions of habitat suitability quantified based on citizen-science data and scientific data from aerial surveys or satellite telemetry (Coxen et al., 2017; Tanner et al., 2020). Similarly to Tiago et al. (2017), we find large interspecific differences in niche overlap, although in our study the proportion of the environmental niche area covered by both the scientific and citizen-science niches is on average around 80% and never lower than half of the total niche. The lower values might be observed because the authors used as their reference scientific dataset atlas data spanning multiple years and therefore with more available observations. Additionally, amphibians and reptiles may be more furtive and difficult to detect for volunteers in comparison to ants. Compared to our findings regarding niche overlap (i.e., in environmental space), the results of SDM predictions (i.e., in geographic space) present more variation. The similarity of response curves and predictor importance between models based on the scientific or the citizen-science dataset is consistent with the conclusions of Tye et al. (2017), who investigated the bias of citizen-science data when modelling the distribution of a rare squirrel species. The variability in correlation between model predictions indicates that they may be more sensitive to species characteristics, however because of the role of chance when selecting predictors and calibrating the models, the absolute values should be interpreted with caution. Overall, the relationship between model predictions suggests that for species with low correlations, scientific and citizen-science models are in agreement on low suitability or unsuitable habitats, but in contrast concerning areas of high suitability.

Broadly speaking, using citizen-science data in SDMs can lead to over-estimating the suitability of urban areas as habitats for ant species. This is reported both for species occurring frequently in urban areas with consistent model predictions between the scientific and the citizen-science models (*L. niger, S. fugax*), and for species occurring in natural and peri-urban areas for which the scientific and citizen-science models favored the respective habitat type (*F. cunicularia, F. fusca*). These results concur with previous studies across a wide range of taxonomic groups demonstrating that citizen-science data are often biased towards highly populated areas, closer to settlements and with denser road networks, which are more accessible to volunteers (Botts et al., 2011; Cretois et al., 2021; Mair & Ruete, 2016). For example, Planillo et al. (2021) studied nightingale distributions in Berlin using SDMs and showed that opportunistic citizen-science data greatly over-predicted the suitability of urban areas compared to more important remote areas. The overestimation of urban habitat suitability can be linked to the results of environmental niche quantifications, as citizen-science niches were located in drier and warmer areas in the environmental space. Urban areas in the canton of Vaud are found at lower elevations in the central plateau, where temperature is higher and precipitations lower. Additionally, a heat island effect is often observed in urban habitats (Rizwan et al., 2008), and studies have shown that urban areas can facilitate the establishment of ant species adapted to drier and warmer environments (Menke et al., 2011).

Species-specific characteristics such as habitat preferences can partially explain the interspecific differences we observe in niche overlap and correlation between model predictions. In general, consistent predictions between the scientific and citizen-science models are produced for species occurring in specific habitat types, such as grasslands (*F. lemani, F. pressilabris*), forests (*F. lugubris/paralugubris*) or rural and urban areas (*L. niger*). On the other hand, generalists occurring in both natural and urban or disturbed habitats show a larger divergence in model predictions, especially more thermophilic species (*F. cunicularia, F. fusca, M. sabuleti*). As a result, the size of the environmental niche can be an important factor determining if citizen-science data are appropriate to model the distribution of a given species. Reductions in niche size can provide useful information about potential increases in species' extinction threat (Breiner et al., 2017), and high-quality results of niche quantification analyses for species with lower niche breadth could be encouraging for the use of citizen-science data in IUCN Red List assessments of vulnerable species. Indeed, both overlap between niches quantified with scientific or citizen-science data and SDM performance have been found to negatively correlate with niche breadth in previous studies (Connor et al., 2018; Tiago et al., 2017). Our findings confirm the second relationship, however the correlation between niche breadth and Spearman's rho between predictions was slightly negative but not significant. This difference may be due to the small number of species studied but can also indicate the greater role of specific functional traits related to species distributions. Traits related to diet and habitat have been shown to influence the over- or under-prediction of bird species diversity using SDMs (Zurell et al., 2016). Species detectability can also affect SDM performance and predictions of species persistence, especially when it varies between habitats (Lahoz-Monfort et al., 2014; Ruiz-Gutiérrez & Zipkin, 2011). For example, in our study, the low correlation between model predictions for the species *T. nylanderi* can be explained by reduced detection in forest habitats by volunteers, as it is small-sized and frequently nesting in plant material such as hollow nuts or acorns (Seifert, 2018).

Despite differences in model performance between species, the AUC indicated fair performance or above for all our models using scientific or citizen-science data. Model performance during evaluation is commonly used to compare SDM results and multiple studies have shown that citizen-science models perform on average similarly to scientific models (Johnston et al., 2020; Tanner et al., 2020; Tye et al., 2017). However, when model evaluation is done internally, for example through cross-validation, the value of performance metrics may remain high even when spatial bias is strong, because this bias is also present in the evaluation data (Beck et al., 2014). For this reason, the AUC values obtained during this study cannot be used to directly compare the models produced with the two datasets and should only be considered as an indication of the quality of model predictions. Ideally, to confirm that citizen-science model results are in good agreement with those based on scientific surveys, an external scientific evaluation dataset should be used (Matutini et al., 2021). Unfortunately, since the number of observations was limited for a lot of species in our scientific data, we could not follow such an approach. Future surveys of ant diversity in the canton of Vaud could potentially provide valuable data to complement and further support our results.

Correction of bias in citizen-science data

The increase in both the niche overlap and the correlation between the predictions of the models using the scientific or the citizen-science data suggests that the correction approach developed in the current study has the potential to reduce the environmental bias present in large citizen-science datasets. The utility of the correction method is further demonstrated when considering that the highest improvement is observed for the species with the lowest values of niche overlap and correlation between predictions. In the case of the species *F. cunicularia* and *F. fusca* for example, there is a clear bias of the citizen-science dataset towards urban areas reflected in the predictions of the citizen-science models that is greatly reduced after the application of environmental correction. On the other hand, the more moderate increase in niche overlap for the species with the largest values is expected as the environmental bias in the citizen-science data is lower. The absence of improvement or decrease in the degree of correlation between the predictions of models with the scientific and corrected citizen-science dataset for species with previously high correlation can be justified in part by the role of random effects during species distribution modelling. Another possible explanation for this result is that environmental variables for modelling were selected using the scientific dataset and one of the corrected subsets of the citizen-science dataset instead of repeating the process for each subset separately. As a result, some models may not have had the optimal predictors, which could have made predictions more similar to those of the scientific dataset.

Compared to alternative methods designed to account for bias in citizen-science data, our approach offers a relatively simple way to correct environmental bias when the density of sampling effort across the study area is unknown. This information is necessary for bias correction methods that rely on using proxies of volunteer sampling effort as predictors in models or to determine the selection of background points, for example by sampling more points in areas with higher effort (Milanesi et al., 2020; Rutten et al., 2019; Ver Hoef et al., 2021). Methods based on integrating observations from multiple species or datasets can handle the presence of bias without explicit information on sampling effort, however they require the application of more complex modelling algorithms, such as point-process models, which may be less accessible to many ecologists (Fithian et al., 2015; Isaac et al., 2020). Another option often preferred when dealing with biased citizen-science data is to reduce spatial bias by thinning, that is removing observations highly clustered in space (Aiello-Lammens et al., 2015). Although spatial thinning in many cases can improve SDM performance, it frequently fails in successful bias correction and can be inappropriate to use for rare species and more generally with datasets of small sample sizes due to the reduction in the number of species' occurrences (Johnston et al., 2020; Steen et al., 2021). Additionally, recent studies showed that the distribution of bias in environmental space plays an important role in the success of spatial thinning approaches (Baker et al., 2022; Kadmon et al., 2004), and indicated that applying environmental filters when thinning observations improves model performance compared to geographical filters (Varela et al., 2014). By subsampling citizen-science observations within environmental clusters, our approach implements environmental thinning and takes advantage of the scientific dataset to estimate differences in sampling effort across environmental conditions.

Although the method developed in the current study has the potential to improve bias correction in citizen-science datasets, there also some limitations that should be considered. Firstly, a prerequisite to apply this technique is the presence of scientific data in the study area for the species of interest, yet these data may be missing for under-studied areas or taxonomic groups (Amano et al., 2016). Secondly, a core assumption of our method is that the reference scientific dataset is unbiased with regards to sampling effort across the study area. Previous studies have shown that structured scientific datasets can also be biased, for example with regards to roads (Tye et al., 2017). In this case, environmental bias correction could lead to replacing the citizen-science data bias with the scientific data bias in model predictions. The random-stratified sampling scheme used to collect the scientific data in the current study ensures that the sampling bias with regards to habitat types, including transportation infrastructures, is limited, however this consideration should be taken into account when applying the method to other datasets. Another limitation is that, when the number of observations in the scientific dataset is low, citizen-science data can bring new information on the distribution of a species (Dickinson et al., 2010), which could be lost during correction. Finally, similarly to spatial thinning, environmental thinning also greatly decreases the number of observations in the citizen-science dataset. Still, models based on few observations of high quality often outperform uncorrected datasets (Varela et al., 2014). A potential solution to the last two limitations would be to apply partial environmental thinning, where the frequency of scientific observations by environmental cluster is used as weights to subsample the citizen-science dataset, instead of sampling strictly the same proportion of points. This approach can both retain some of the information of the citizen-science dataset concerning the spatial distribution of species' occurrences and ensure that a higher number of observations is kept.

Integrated models and future perspectives

In this study, we combined scientific and corrected citizen-science data to model the environmental niches and distributions of 15 ant species in the canton of Vaud from a wide range of habitat types, such as species associated with urban areas (*L. niger*), mountainous and alpine grasslands (*F. pressilabris*) and coniferous or mixed forests (*F. lugubris/paralugubris*). The predictions of the integrated models offer an overview of ant diversity in the study area and highlight important sites for ants, including the high elevation areas of the Jura mountains and the western Swiss Pre-Alps. Ant species are protected in the canton of Vaud (Avril et al., 2019), and there is great interest in their conservation, especially in the case of wood ant colonies (*Formica s.str* sub-genus) in national parcs (Parc Jura vaudois, 2022). This is not surprising considering that ants participate in various ecosystem services, such as seed dispersal, pollination, and regulation of pest populations (Del Toro et al., 2012). In the soil, ants act as ecosystem engineers and can modify its physical and chemical properties during nest excavation (Folgarait, 1998). Citizen-science data can be invaluable to study this important taxonomic group, allowing to model the distributions of a larger number of species compared to using scientific data only. Future studies could expand our modelling strategy to produce maps of ant species richness in the canton of Vaud, in order to guide subsequent sampling campaigns or for conservation planning purposes.

The consistency in environmental niches and SDM results between the scientific and the citizen-science dataset also underlines that citizen-science data have many promising applications in future ecological and conservation studies. With regards to the Opération Fourmis data, several opportunities for further research exist. For example, the degree of niche overlap between closely related ant species can be used to study if pairs of species with higher hybridization rates have more similar environmental niches (Lavanchy, 2022). SDM results can also be used to predict the spread of invasive ant species in canton of Vaud, for example by combining them with a model simulating dispersal or migration (e.g. MigClim, Engler *et al.*, 2012), or to study the impacts of climate and land-use change on ant distributions (Bertelsmeier et al., 2016; Bujan et al., 2021; Del Toro et al., 2015).

In conclusion, our results highlight how citizen-science data can reliably be used to model species environmental niches and distributions. Nevertheless, researchers should be mindful of the characteristics of the species modelled, such as habitat preferences and detectability, in conjunction with the distribution of environmental conditions in their study area. Correcting the environmental bias present in the citizen-science data can be a simple and efficient solution to avoid overestimating the importance of unsuitable habitats.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Olivier Brönnimann for his continuous support and guidance during this project. I am grateful to Antoine Guisan for his recommendations and constructive suggestions along with the anonymous referee. I would also like to thank Tanja Schwander for her help with the additional ant data collection and sample identification and her advice. I want to thank Anne Freitag for her help with ant species' morphological identifications, as well as Marjorie Labédan for her technical assistance with ant genetic identifications. I would like to thank Timothy Szewczyk for providing information about the structure of the datasets and the sampling design. Special thanks to Antoine Adde and Flavien Collart for their help with selecting model predictors and ESMs respectively, and all the members of the ecospat group for their productive suggestions. Finally, this project would not have been possible without the contributions of the volunteers and staff of the "Opération Fourmis" project.

REFERENCES

Adde, A., Rey, P.-L., Brun, P., Kulling, N., Fopp, F., Altermatt, F., Broennimann, O., Lehmann, A., Petitpierre, B., Zimmermann, N., Pellissier, L., & Guisan, A. (2022). *A high-performance computing pipeline for Nested Species Distribution Modelling*. In review for Ecography.

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*, 541–545. https://doi.org/10.1111/ecog.01132

Amano, T., Lamming, J. D. L., & Sutherland, W. J. (2016). Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. *BioScience*, *66*(5), 393–400. https://doi.org/10.1093/biosci/biw022

Anderson, R. P., Araújo, M. B., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, A. T., & Soberón, J. M. (2020). Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Frontiers of Biogeography*, *12*(3), e47839.

Avril, A., Depraz, A., & Schwander, T. (2019). Opération Fourmis, le premier recensement participatif des fourmis vaudoises: Contexte, méthodologie et bilan préliminaire. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *98*, 109–120. https://doi.org/10.5169/SEALS-846640

Baker, D. J., Maclean, I. M. D., Goodall, M., & Gaston, K. J. (2022). Correlations between spatial sampling biases and environmental niches affect species distribution models. *Global Ecology and Biogeography*, *31*(6), 1038–1050. https://doi.org/10.1111/geb.13491

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bertelsmeier, C., Blight, O., & Courchamp, F. (2016). Invasions of ants (Hymenoptera: Formicidae) in light of global climate change. *Myrmecological News*, *22*, 25–42.

Botts, E. A., Erasmus, B. F. N., & Alexander, G. J. (2011). Geographic sampling bias in the South African Frog Atlas Project: Implications for conservation planning. *Biodiversity and Conservation*, *20*(1), 119–139. https://doi.org/10.1007/s10531-010-9950-6

Breiner, F. T., Guisan, A., Bergamini, A., & Nobis, M. P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, *6*(10), 1210–1218. https://doi.org/10.1111/2041-210X.12403

Breiner, F. T., Guisan, A., Nobis, M. P., & Bergamini, A. (2017). Including environmental niche information to improve IUCN Red List assessments. *Diversity and Distributions*, *23*(5), 484–495. https://doi.org/10.1111/ddi.12545

Breiner, F. T., Nobis, M. P., Bergamini, A., & Guisan, A. (2018). Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution*, *9*(4), 802–808. https://doi.org/10.1111/2041-210X.12957

Broennimann, O. (2018). CHclim25: A high spatial and temporal resolution climate dataset for Switzerland. *Ecospat Laboratory, University of Lausanne: Lausanne, Switzerland*.

Broennimann, O., Di Cola, V., & Guisan, A. (2022). *Ecospat: Spatial Ecology Miscellaneous Methods. R package version 3.4* (v. 3.4) [R package]. http://www.unil.ch/ecospat/home/menuguid/ecospat-resources/tools.html

Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., Thuiller, W., Fortin, M.-J., Randin, C., Zimmermann, N. E., Graham, C. H., & Guisan, A. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, *21*(4), 481–497. https://doi.org/10.1111/j.1466-8238.2011.00698.x

Bujan, J., Charavel, E., Bates, O. K., Gippet, J. M. W., Darras, H., Lebas, C., & Bertelsmeier, C. (2021). Increased acclimation ability accompanies a thermal niche shift of a recent invasion. *Journal of Animal Ecology*, *90*(2), 483–491. https://doi.org/10.1111/1365-2656.13381

Callaghan, C. T., Poore, A. G. B., Mesaglio, T., Moles, A. T., Nakagawa, S., Roberts, C., Rowley, J. J. L., VergÉs, A., Wilshire, J. H., & Cornwell, W. K. (2021). Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience*, *71*(1), 55–63. https://doi.org/10.1093/biosci/biaa131

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

Chauvier, Y., Zimmermann, N. E., Poggiato, G., Bystrova, D., Brun, P., & Thuiller, W. (2021). Novel methods to correct for observer and sampling bias in presence-only species distribution models. *Global Ecology and Biogeography*, *30*(11), 2312–2325. https://doi.org/10.1111/geb.13383

Chen, Y. (2008). Global potential distribution of an invasive species, the yellow crazy ant (*Anoplolepis gracilipes*) under climate change. *Integrative Zoology*, *3*(3), 166–175. https://doi.org/10.1111/j.1749-4877.2008.00095.x

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F., & Liu, J. (2018). Effects of grain size and niche breadth on species distribution modeling. *Ecography*, *41*(8), 1270–1282. https://doi.org/10.1111/ecog.03416

Cooper, C. B., Shirk, J., & Zuckerberg, B. (2014). The Invisible Prevalence of Citizen Science in Global Research: Migratory Birds and Climate Change. *PLOS ONE*, *9*(9), e106508. https://doi.org/10.1371/journal.pone.0106508

Coxen, C. L., Frey, J. K., Carleton, S. A., & Collins, D. P. (2017). Species distribution models for a migratory bird based on citizen science and satellite tracking data. *Global Ecology and Conservation*, *11*, 298–311. https://doi.org/10.1016/j.gecco.2017.08.001

Cretois, B., Simmonds, E. G., Linnell, J. D. C., van Moorter, B., Rolandsen, C. M., Solberg, E. J., Strand, O., Gundersen, V., Roer, O., & Rød, J. K. (2021). Identifying and correcting spatial bias in opportunistic citizen science data for wild ungulates in Norway. *Ecology and Evolution*, *11*(21), 15191–15204. https://doi.org/10.1002/ece3.8200

Del Toro, I., Ribbons, R. R., & Pelini, S. L. (2012). The little things that run the world revisited: A review of ant-mediated ecosystem services and disservices (Hymenoptera: Formicidae). *Myrmecological News*, *17*, 133–146.

Del Toro, I., Silva, R. R., & Ellison, A. M. (2015). Predicted impacts of climatic change on ant functional diversity and distributions in eastern North American forests. *Diversity and Distributions*, *21*(7), 781–791. https://doi.org/10.1111/ddi.12331

Descombes, P., Walthert, L., Baltensweiler, A., Meuli, R. G., Karger, D. N., Ginzler, C., Zurell, D., & Zimmermann, N. E. (2020). Spatial modelling of ecological indicator values improves predictions of plant distributions in complex landscapes. *Ecography*, *43*(10), 1448–1463. https://doi.org/10.1111/ecog.05117

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., d'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., & Dubuis, A. (2017). ecospat: An R package to support spatial analyses and modeling of species niches and distributions. *Ecography*, *40*(6), 774–787.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Engler, R., Hordijk, W., & Guisan, A. (2012). The MIGCLIM R package – seamless integration of dispersal constraints into projections of species distribution models. *Ecography*, *35*(10), 872–878. https://doi.org/10.1111/j.1600-0587.2012.07608.x

Feldman, M. J., Imbeau, L., Marchand, P., Mazerolle, M. J., Darveau, M., & Fenton, N. J. (2021). Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PLOS ONE*, *16*(3), e0234587. https://doi.org/10.1371/journal.pone.0234587

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), 424–438. https://doi.org/10.1111/2041-210X.12242

Fitzpatrick, M. C., Gotelli, N. J., & Ellison, A. M. (2013). MaxEnt versus MaxLike: Empirical comparisons with ant species distributions. *Ecosphere*, *4*(5), 1–15. https://doi.org/10.1890/ES13-00066.1

Fletcher Jr., R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*(6), e02710. https://doi.org/10.1002/ecy.2710

Folgarait, P. J. (1998). Ant biodiversity and its relationship to ecosystem functioning: A review. *Biodiversity & Conservation*, *7*(9), 1221–1244. https://doi.org/10.1023/A:1008891901953

Freitag, A., Schwander, T., Broennimann, O., & Dèpraz, A. (2020). Opération Fourmis, les résultats du premier recensement participatif des espèces de fourmis vaudoises. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *99*, 13–28. https://doi.org/10.5169/SEALS-917230

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gaiji, S., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., & Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodiversity Informatics*, *8*(2), 94-172. https://doi.org/10.17161/bi.v8i2.4124

Gardiner, M. M., Allee, L. L., Brown, P. M., Losey, J. E., Roy, H. E., & Smyth, R. R. (2012). Lessons from lady beetles: Accuracy of monitoring data from US and UK citizen-science programs. *Frontiers in Ecology and the Environment*, *10*(9), 471–476. https://doi.org/10.1890/110185

GBIF: The Global Biodiversity Information Facility. (2022, December 6). *What is GBIF?* https://www.gbif.org/what-is-gbif

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, *19*(9), 497–503.

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., & Kueffer, C. (2014). Unifying niche shift studies: Insights from biological invasions. *Trends in Ecology & Evolution*, *29*(5), 260–269.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge University Press.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Hartley, S., Harris, R., & Lester, P. J. (2006). Quantifying uncertainty in the potential distribution of an invasive species: Climate and the Argentine ant: Quantifying uncertainty in range map models. *Ecology Letters*, *9*(9), 1068–1079. https://doi.org/10.1111/j.1461-0248.2006.00954.x

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., Lamigueiro, O. P., Bevan, A., Racine, E. B., & Shortridge, A. (2022). *Raster: Geographic Data Analysis and Modeling. R package version 3.5-15* (3.5-15). https://CRAN.R-project.org/package=raster

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*(2), 130–137. https://doi.org/10.1016/j.tree.2011.11.006

Hochmair, H. H., Scheffrahn, R. H., Basille, M., & Boone, M. (2020). Evaluating the data quality of iNaturalist termite records. *PLOS ONE*, *15*(5), e0226534. https://doi.org/10.1371/journal.pone.0226534

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Johnston, A., Matechou, E., & Dennis, E. B. (2022). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103–116. https://doi.org/10.1111/2041-210X.13834

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*, 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models. *Ecological Applications*, *14*(2), 401–413. https://doi.org/10.1890/02-5364

Kremen, C., Ullman, K. S., & Thorp, R. W. (2011). Evaluating the Quality of Citizen-Scientist Data on Pollinator Communities. *Conservation Biology*, *25*(3), 607–617. https://doi.org/10.1111/j.1523-1739.2011.01657.x

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Wintle, B. A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, *23*(4), 504–515. https://doi.org/10.1111/geb.12138

Larson, E. R., Graham, B. M., Achury, R., Coon, J. J., Daniels, M. K., Gambrell, D. K., Jonasen, K. L., King, G. D., LaRacuente, N., Perrin-Stowe, T. I., Reed, E. M., Rice, C. J., Ruzi, S. A., Thairu, M. W., Wilson, J. C., & Suarez, A. V. (2020). From eDNA to citizen science: Emerging tools for the early detection of invasive species. *Frontiers in Ecology and the Environment*, *18*(4), 194–202. https://doi.org/10.1002/fee.2162

Lavanchy, G. (2022). *Evolutionary causes and consequences of hybridization in insects* [Doctoral Thesis, University of Lausanne]. https://serval.unil.ch/en/notice/serval:BIB_FB45E1DEBEDB

Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J., & Guisan, A. (2010). Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. *Biological Conservation*, *143*(11), 2647–2657.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *Cluster: Cluster Analysis Basics and Extensions. R package version 2.1. 2*.

Mair, L., & Ruete, A. (2016). Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. *PLOS ONE*, *11*(1), e0147796. https://doi.org/10.1371/journal.pone.0147796

Matutini, F., Baudry, J., Pain, G., Sineau, M., & Pithon, J. (2021). How citizen science could improve species distribution models and their independent assessment. *Ecology and Evolution*, *11*(7), 3028–3039. https://doi.org/10.1002/ece3.7210

McKinley, D. C., Miller-Rushing, A. J., Ballard, H. L., Bonney, R., Brown, H., Cook-Patton, S. C., Evans, D. M., French, R. A., Parrish, J. K., Phillips, T. B., Ryan, S. F., Shanley, L. A., Shirk, J. L., Stepenuck, K. F., Weltzin, J. F., Wiggins, A., Boyle, O. D., Briggs, R. D., Chapin, S. F., … Soukup, M. A. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, *208*, 15–28. https://doi.org/10.1016/j.biocon.2016.05.015

Menke, S. B., Guénard, B., Sexton, J. O., Weiser, M. D., Dunn, R. R., & Silverman, J. (2011). Urban areas may serve as habitat and corridors for dry-adapted, heat tolerant species; an example from ants. *Urban Ecosystems*, *14*(2), 135–163. https://doi.org/10.1007/s11252-010-0150-7

Milanesi, P., Mori, E., & Menchetti, M. (2020). Observer-oriented approach improves species distribution models from citizen science data. *Ecology and Evolution*, *10*(21), 12104–12114. https://doi.org/10.1002/ece3.6832

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Parc Jura vaudois. (2022, December 20). *Chronique d'une fourmi des bois*. https://parcjuravaudois.ch/chronique-dune-fourmi

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Planillo, A., Fiechter, L., Sturm, U., Voigt-Heucke, S., & Kramer-Schadt, S. (2021). Citizen science data for urban planning: Comparing different sampling schemes for modelling urban bird distribution. *Landscape and Urban Planning, 211,* 104098. https://doi.org/10.1016/j.landurbplan.2021.104098

Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D., & Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *PLOS ONE*, *12*(4), e0172579. https://doi.org/10.1371/journal.pone.0172579

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rizwan, A. M., Dennis, L. Y., & Chunho, L. (2008). A review on the generation, determination and mitigation of Urban Heat Island. *Journal of Environmental Sciences*, *20*(1), 120–128.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruiz-Gutiérrez, V., & Zipkin, E. F. (2011). Detection biases yield misleading patterns of species persistence and colonization in fragmented landscapes. *Ecosphere*, *2*(5), art61. https://doi.org/10.1890/ES10-00207.1

Rutten, A., Casaer, J., Swinnen, K. R. R., Herremans, M., & Leirs, H. (2019). Future distribution of wild boar in a highly anthropogenic landscape: Models combining hunting bag and citizen science data. *Ecological Modelling*, *411*, 108804. https://doi.org/10.1016/j.ecolmodel.2019.108804

Seifert, B. (2018). *The ants of central and north Europe*. Lutra Verlags-und Vertriebsgesellschaft.

Steen, V. A., Tingley, M. W., Paton, P. W. C., & Elphick, C. S. (2021). Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution*, *12*(2), 216–226. https://doi.org/10.1111/2041-210X.13525

Szewczyk, T. M., Bertelsmeier, C., Schwander, T., & Anne, C. (2022). *Leveraging citizen science to assess richness, diversity, and abundance*. [Preprint]

Tanner, A. M., Tanner, E. P., Papeş, M., Fuhlendorf, S. D., Elmore, R. D., & Davis, C. A. (2020). Using aerial surveys and citizen science to create species distribution models for an imperiled grouse. *Biodiversity and Conservation*, *29*(3), 967–986. https://doi.org/10.1007/s10531-019-01921-6

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M. A., & Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Tiago, P., Pereira, H. M., & Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions. *Basic and Applied Ecology*, *20*, 75–85. https://doi.org/10.1016/j.baae.2017.04.001

Tulloch, A. I. T., Mustin, K., Possingham, H. P., Szabo, J. K., & Wilson, K. A. (2013). To boldly go where no volunteer has gone before: Predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, *19*(4), 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x

Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., & Martin, T. G. (2013). Realising the full potential of citizen science monitoring programs. *Biological Conservation*, *165*, 128–138. https://doi.org/10.1016/j.biocon.2013.05.025

Tye, C. A., McCleery, R. A., Fletcher Jr, R. J., Greene, D. U., & Butryn, R. S. (2017). Evaluating citizen vs. Professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, *54*(2), 628–637. https://doi.org/10.1111/1365-2664.12682

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Ver Hoef, J. M., Johnson, D., Angliss, R., & Higham, M. (2021). Species density models from opportunistic citizen science data. *Methods in Ecology and Evolution*, *12*(10), 1911–1925. https://doi.org/10.1111/2041-210X.13679

Ward, D. F. (2007). Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions*, *9*(6), 723–735. https://doi.org/10.1007/s10530-006-9072-y

Zurell, D., Zimmermann, N. E., Sattler, T., Nobis, M. P., & Schröder, B. (2016). Effects of functional traits on the prediction accuracy of species richness models. *Diversity and Distributions*, *22*(8), 905–917. https://doi.org/10.1111/ddi.12450

SUPPLEMENTARY METHODS

Genetic identification of ant species

The additional ant samples collected during spring-summer 2022 in the context of this project were identified genetically at the species level with the following protocol. For each sampled colony of the *Temnothorax* and *Tetramorium* genera, as well as *Myrmica* species that could not be identified morphologically as *Myrmica rubra* or *Myrmica ruginodis*, a single worker was chosen for genetic identification based on the mitochondrial COI gene. After removing the abdomen, each individual was placed in a vial in liquid azote for 5 minutes and afterwards ground at 6200 rpm for 30 seconds. 180 µl of buffer ATL and 20 µl were added to each tube, and the samples were left to digest overnight at 56 ^{o}C and 30 rpm. DNA extraction was performed using the QIAGEN DNeasy Blood & Tissue kit with the following change: 100 µl of EB buffer were used for DNA elution. PCR was performed using the LCO and HCO primers, following the protocol in Lavanchy (2022), and DNA concentration was assessed through gel electrophoresis. Samples were sent to Microsynth AG, Switzerland, for Sanger sequencing. Sequencing trace files were corrected and trimmed with the MEGA-X software (Kumar et al., 2018) and samples with ambiguous results were excluded from further analyses. For the *Tetramorium* and *Myrmica* individuals, COI sequence reads were aligned to previously analyzed *Tetramorium* and *Myrmica* sequences respectively from the 2019 Opération Fourmis campaign using the MUSCLE algorithm with default settings. *Temnothorax* samples were excluded because of the poor quality of sequencing results. Maximum likelihood phylogenies for each genus were built in MEGA-X using the GTR + I + Γ model of DNA evolution with 100 bootstrap repetitions and fast SRP branch swapping. To identify the species, the clustering of the new samples with the already identified 2019 Opération Fourmis samples in the phylogeny was investigated. The bootstrap values were used to determine if the identification could be done confidently. When the result was ambiguous because of clustering with Opération Fourmis samples belonging to multiple species or low bootstrap values, the sample was considered as unidentified and not included in the models.

Predictor selection

To determine the most appropriate method for the selection of modelling environmental predictors, three alternative approaches were assessed using the data from the 2019 Opération Fourmis project. First, a PCA was carried out using the values of the environmental variables at the coordinates of the species observations for both datasets, and the first principal components explaining 95% of the variance were retained for each species. Secondly, a penalized regression using Elastic Net GLM was performed using the glmnet package v.4.1-2 (Friedman et al., 2010), and the number of retained variables at the end of the analysis was assessed. For each species the occurrences in both datasets were combined with a set of 10,000 randomly selected background points across the study area, following Barbet-Massin et al. (2012). Weights were applied to the observations to control for prevalence and give equal weight to presences and background points. All the predictors were added as both linear and quadratic terms to the regression, as especially for climatic variables quadratic terms represent better the response curve of a species. Finally, the nsdm package was used as described in the main text. The number of predictors selected by each method is given in Table S3. Although both the PCA and the nsdm method yield comparable results and allow for ranking of the predictors based on their importance, nsdm provides a ranking of the original predictors, which are more easily interpretable than the principal components obtained through the PCA.

SUPPLEMENTARY MATERIALS

Species occurrence data

Table S1. Number of occurrences for each of the selected species in i) the scientific dataset, ii) the citizen-science dataset, iii) the subsamples of the citizen-science dataset after correction of the environmental bias.

Species	Scientific Data	Citizen-science Data	Corrected C-S Data
Formica cunicularia	30	386	15
Formica fusca	33	125	39
Formica lemani	115	140	42
Formica lugubris/paralugubris	87	449	314
Formica pressilabris	181	79	73
Lasius flavus	194	306	85
Lasius niger	241	1300	312
Myrmica rubra	26	86	36
Myrmica ruginodis	133	244	40
Myrmica sabuleti	42	116	35
Myrmica scabrinodis	82	33	14
Solenopsis fugax	44	100	39
Temnothorax nylanderi - GR	73	66	25
Tetramorium caespitum	46	286	78
Tetramorium impurum	24	53	18

Environmental Niche Quantifications

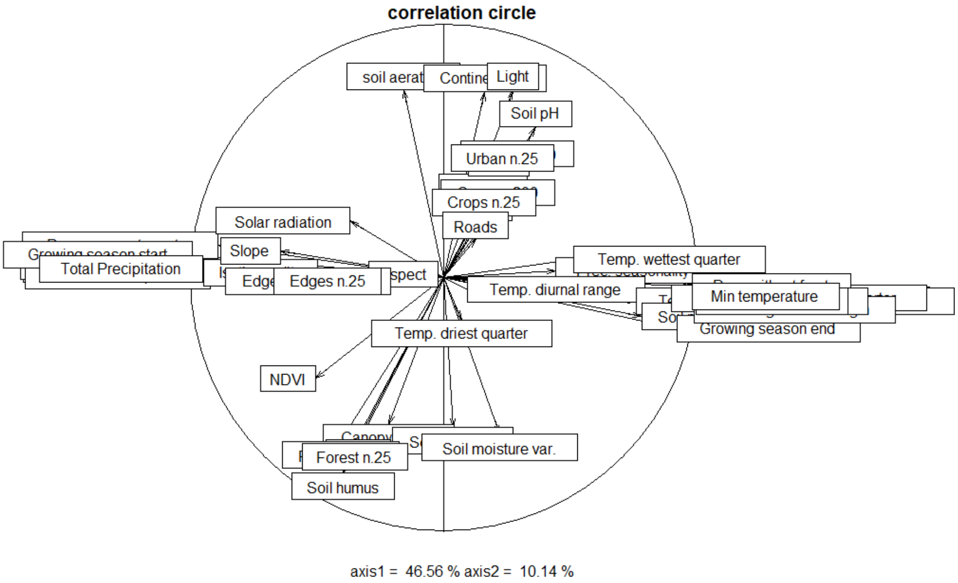


Figure S1. Correlation circle using the first two axes of the PCA over the 59 environmental predictors chosen in the study area.

Table S2. Results of environmental niche modelling for each species using the scientific and the citizen-science dataset. Statistically significant values are given in bold.

Species	**Overlap D**	**Niche area covered by the scientific dataset (unfilling)**	**Niche area covered by both datasets (stability)**	**Niche area covered by the citizen-science dataset (expansion)**
Formica cunicularia	**0.344**	0.133	0.7	0.3
Formica fusca	**0.38**	0.216	0.517	0.483
Formica lemani	**0.689**	**0.025**	**0.84**	**0.16**
Formica lugubris/paralugubris	**0.719**	**0.007**	**0.811**	**0.189**
Formica pressilabris	**0.769**	**0.033**	**0.946**	**0.054**
Lasius flavus	**0.704**	**0.011**	**0.925**	**0.075**
Lasius niger	**0.595**	**0.003**	**0.906**	**0.094**
Myrmica rubra	**0.639**	0.113	0.782	0.218
Myrmica ruginodis	**0.809**	**0.027**	**0.936**	**0.064**
Myrmica sabuleti	**0.562**	0.022	0.742	0.258
Myrmica scabrinodis	**0.668**	**0.026**	**0.794**	**0.206**
Solenopsis fugax	**0.7**	**0.09**	0.897	0.103
Temnothorax nylanderi - GR	**0.473**	0.237	0.737	0.263
Tetramorium caespitum	**0.755**	**0.03**	**0.928**	**0.072**
Tetramorium impurum	**0.746**	**0**	**0.785**	**0.215**

Predictor selection

Table S3. Number of environmental predictors selected for modelling using each predictor selection method: i) Elastic Net GLM. The number of predictors based on the lambda with the least square error is given within the parenthesis and the number of predictors for the value of lambda at 1 standard deviation of the least square error is given outside the parenthesis. Because predictors are added as both linear and quadratic terms, the maximum number of potential predictors for this method is 118. ii) Principal Component Analysis, iii) NSDM package.*

Species	**Elastic Net GLM**	**PCA**	**NSDM**
Formica cunicularia	63 (63)	17	20
Formica fusca	42 (48)	16	15
Formica lemani	12 (58)	16	17
Formica lugubris/paralugubris	35 (57)	17	15
Formica pressilabris	43 (46)	15	19
Lasius flavus	40 (64)	17	19
Lasius niger	49 (101)	19	19
Myrmica rubra	77 (88)	16	16
Myrmica ruginodis	41 (82)	16	17
Myrmica sabuleti	33 (49)	15	18
Myrmica scabrinodis	21 (26)	15	14
Solenopsis fugax	37 (45)	17	17
Temnothorax nylanderi - GR	118 (118)	15	19
Tetramorium caespitum	53 (77)	18	18

Tetramorium impurum	14 (15)	14	16

Table S4. Predictor variables selected for species distribution modelling by i) pooling the scientific dataset and the citizen-science dataset for each species before the correction, ii) pooling the scientific dataset and a subset of the citizen-science dataset corrected for environmental bias.

Species	**Predictors before correction**	**Predictors after correction**
Formica cunicularia	Urban area % within 200 m, Light EIV, Continentality EIV, Maximum consecutive days without frost, NDVI, Soil nutrients EIV, Forest area %, Slope	Edge area % within 200 m, NDVI, Canopy height, Duration of daily sunshine, Aspect, Aridity index, Mean diurnal range Urban area % within 25 m
Formica fusca	Annual precipitation, Soil moisture EIV, Edge area % within 200 m, Forest area % within 25 m, Urban area %, NDVI, Soil humus EIV, Aspect	Edge area % within 200 m, Soil nutrients EIV, Continentality EIV, Urban area % within 25 m, Temperature annual range, Mean temperature of driest quarter, Maximum consecutive days without frost, Forest area % within 25 m
Formica lemani	Soil nutrients EIV, Evapotranspiration, Temperature annual range, Forest area % within 25 m, Duration of daily sunshine, Soil moisture EIV, Perimeter of buildings, Aspect	Soil nutrients EIV, Temperature annual range, NDVI, Evapotranspiration, Forest area % within 25 m, Edge area % within 200 m, Isothermality, Continentality EIV
Formica lugubris/paralugubris	Temperature seasonality, Slope, Aspect, Edge area % within 25 m, Soil humus EIV, Edge area % within 200 m, Canopy height, Mean temperature of driest quarter	Temperature annual range, Soil nutrients EIV, Evapotranspiration, Mean temperature of driest quarter, Slope, Aspect, NDVI, Edge area % within 200 m
Formica pressilabris	Temperature annual range, Canopy height, Precipitation of warmest quarter, Mean temperature of wettest quarter, Soil aeration EIV, Isothermality, Slope, Light EIV	Temperature annual range, Canopy height, Mean temperature of wettest quarter, Precipitation of warmest quarter, Soil aeration EIV, Light EIV, Slope, Urban area % within 200 m
Lasius flavus	Continentality EIV, Urban area % within 25 m, NDVI,	Precipitation of warmest quarter, Continentality EIV,

	Edge area % within 200 m, Slope, Aspect, Duration of daily sunshine, Soil humus EIV	Edge area % within 200 m, Mean temperature of driest quarter, Forest area %, NDVI, Aspect, Length of roads
Lasius niger	Urban area % within 25 m, Maximum consecutive days without frost, Edge area % within 200 m, Isothermality, Perimeter of buildings, Aspect, Slope, Temperature annual range	Maximum consecutive days without frost, Slope, Urban area % within 200 m, Soil nutrients EIV, Edge area % within 200 m, Soil humus EIV, Light EIV, NDVI
Myrmica rubra	Edge area % within 200 m, Soil moisture variation EIV, Urban area % within 25 m, Edge area %, Soil humus EIV, Permanent agricultural area % within 200 m, NDVI, Mean diurnal range	Soil moisture variation, NDVI, Edge area % within 200 m, Soil humus EIV, Duration of daily sunshine, Urban area % within 25 m, Edge area %, Isothermality
Myrmica ruginodis	Edge area % within 200 m, Forest area % within 25 m, Precipitation of wettest month, NDVI, Precipitation seasonality, Edge area %, Continentality EIV, Soil humus EIV	Forest area % within 25 m, NDVI, Edge area % within 200 m, Mean temperature of driest quarter, Aridity index, Slope Edge area % within 25 m, Urban area % within 200 m
Myrmica sabuleti	Soil moisture EIV, Soil nutrients EIV, Edge area % within 200 m, Canopy height, Mean temperature of driest quarter, Urban area % within 200 m, Permanent agricultural area % within 200 m, NDVI	Continentality EIV, Mean temperature of wettest quarter, Edge area % within 25 m, Permanent agricultural area % within 25 m, Evapotranspiration Forest area % within 25 m, Aspect, Duration of daily sunshine
Myrmica scabrinodis	Canopy height, Edge area % within 200 m, Temperature seasonality, Slope, Mean temperature of wettest quarter, Urban area % within 200 m, Soil humus EIV, Permanent agricultural area % within 200 m	Temperature seasonality, Canopy height, Edge area % within 200 m, Slope, Urban area % within 200 m, Mean temperature of wettest quarter, Permanent agricultural area % within 200 m, NDVI
Solenopsis fugax	Urban area % within 200 m, Continentality EIV,	Continentality EIV, Urban area % within 200 m, Aridity index,

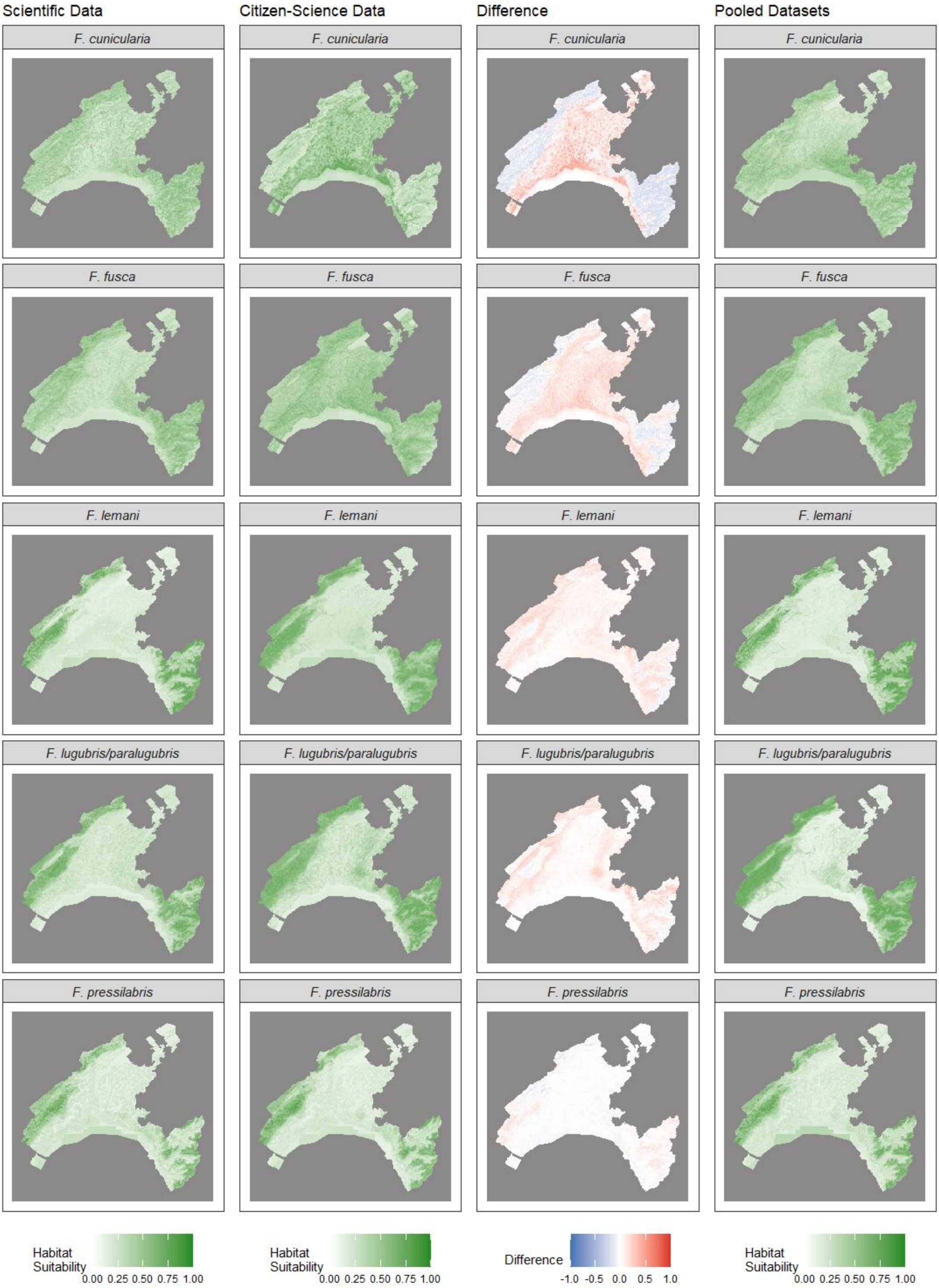
	Permanent agricultural area % within 200 m, Soil moisture variation EIV, Aridity index, NDVI, Mean temperature of driest quarter, Aspect	Light EIV, Edge area % within 25 m, Soil humus EIV, Duration of daily sunshine, Aspect
Temnothorax nylanderi* - *GR	Canopy height, Precipitation seasonality, Edge area % within 200 m, Maximum consecutive days without frost, Soil pH EIV, Continentality EIV, Edge area %, NDVI	Canopy height, Continentality EIV, Duration of daily sunshine, NDVI, Growing season start, Aspect, Edge area %, Precipitation seasonality
Tetramorium caespitum	Urban area % within 200 m, Continentality EIV, Permanent agricultural area % within 200 m, Light EIV, Soil nutrients EIV, Soil moisture variation EIV, Maximum consecutive days without frost, Aspect	Continentality EIV, Aspect, Soil moisture variation EIV, Urban area % within 200 m, Precipitation seasonality, Permanent agricultural area % within 200 m, Maximum consecutive days without frost, Mean temperature of driest quarter
Tetramorium impurum	Soil moisture EIV, Temperature Seasonality, Forest area % within 25 m, Mean temperature of wettest quarter, Aspect, NDVI, Slope, Edge area % within 200 m	Precipitation of coldest quarter, NDVI, Canopy height, Evapotranspiration, Soil nutrients EIV, Continentality EIV, Edge area % within 200 m, Precipitation seasonality

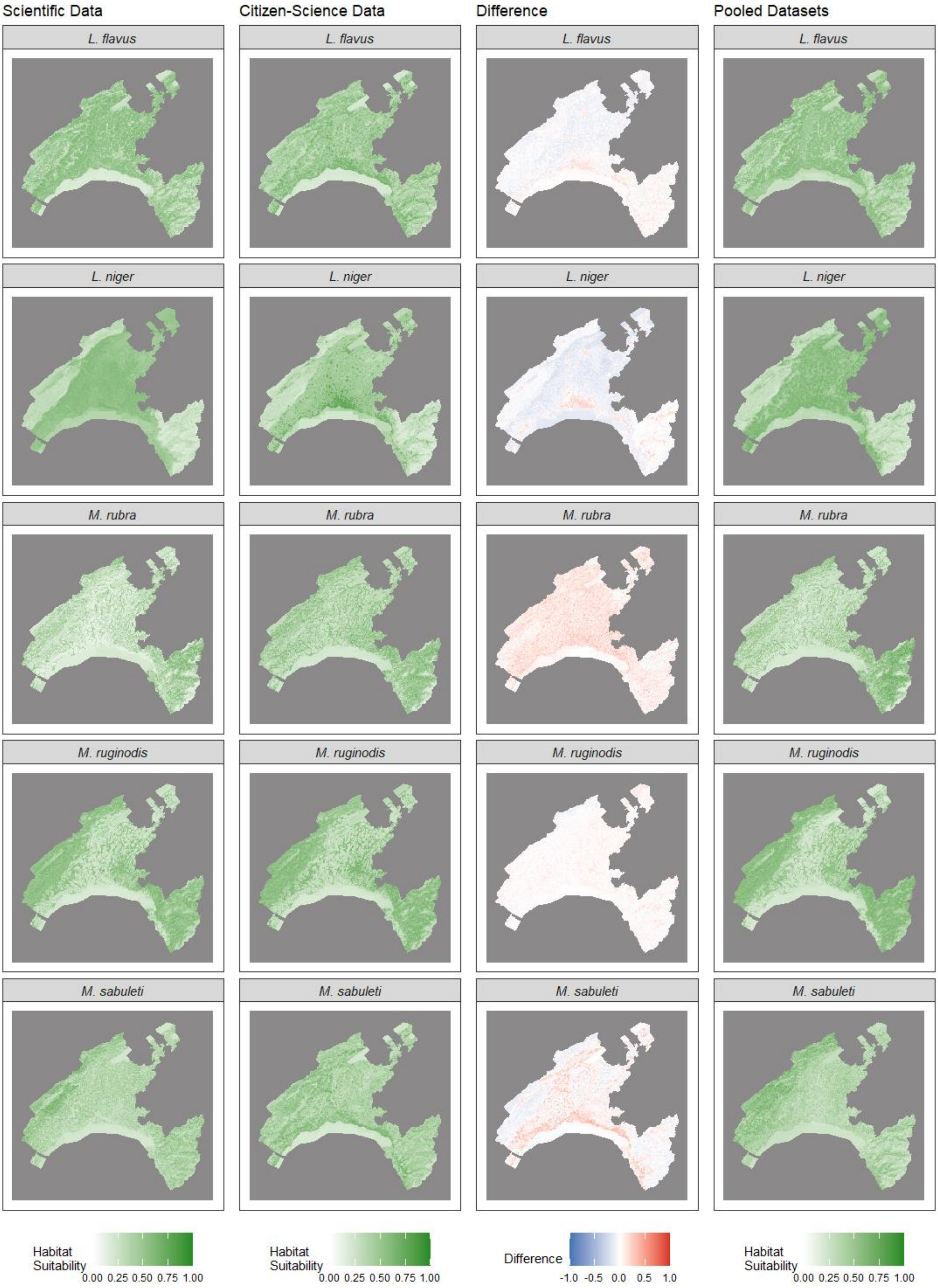
Species Distribution Modelling

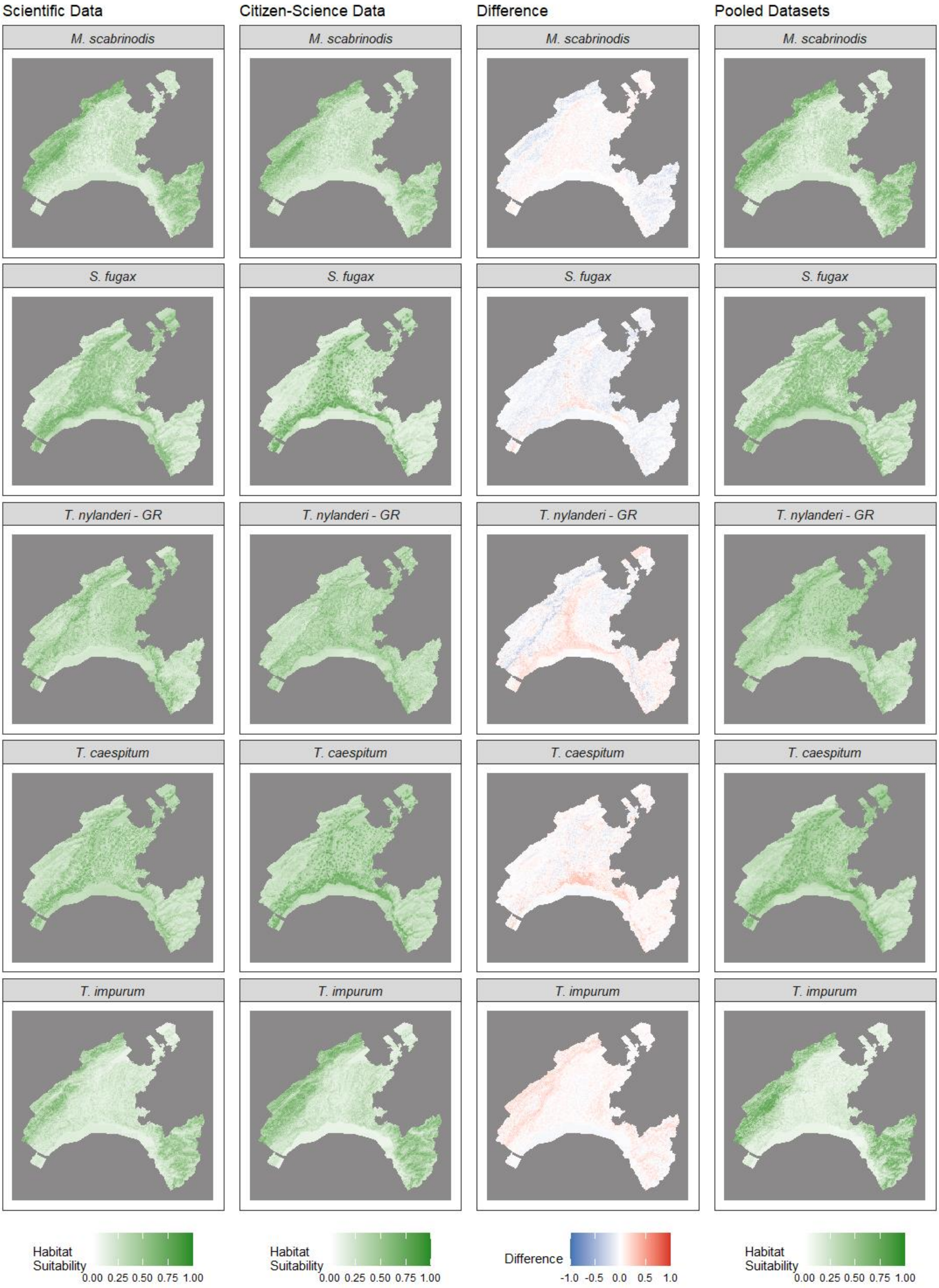
Table S5. Evaluation metrics for each of the ESM species distribution models. For each species a model was created using i) the scientific dataset only, ii) the citizen-science dataset only, iii) a pooled dataset of the scientific data and a subset of the citizen-science data corrected for environmental bias, iv) a corrected subset of the citizen-science dataset if enough occurrences were available. In the last case the mean AUC, MaxTSS and Boyce metrics were calculated from 10 models using 10 different randomly selected subsets respectively.

Species	Model	AUC	MaxTSS	Boyce
Formica cunicularia	Scientific	0.78	0.429	0.942
Formica cunicularia	Citizen-Science	0.874	0.594	0.984
Formica cunicularia	Pooled	0.84	0.58	0.968
Formica fusca	Scientific	0.799	0.483	0.874
Formica fusca	Citizen-Science	0.805	0.487	0.932
Formica fusca	Pooled	0.811	0.473	0.969
Formica fusca	Corrected C-S	0.911	0.697	0.98

Formica lemani	Scientific	0.95	0.776	0.958
Formica lemani	Citizen-Science	0.884	0.643	0.992
Formica lemani	Pooled	0.933	0.746	0.979
Formica lemani	Corrected C-S	0.927	0.745	0.951
Formica lugubris/paralugubris	Scientific	0.911	0.72	0.983
Formica lugubris/paralugubris	Citizen-Science	0.864	0.599	0.994
Formica lugubris/paralugubris	Pooled	0.874	0.648	0.986
Formica lugubris/paralugubris	Corrected C-S	0.863	0.623	0.976
Formica pressilabris	Scientific	0.984	0.908	0.994
Formica pressilabris	Citizen-Science	0.968	0.829	0.985
Formica pressilabris	Pooled	0.978	0.87	0.994
Formica pressilabris	Corrected C-S	0.969	0.838	0.984
Lasius flavus	Scientific	0.788	0.43	0.959
Lasius flavus	Citizen-Science	0.845	0.561	0.986
Lasius flavus	Pooled	0.805	0.472	0.998
Lasius flavus	Corrected C-S	0.838	0.516	0.99
Lasius niger	Scientific	0.82	0.512	0.871
Lasius niger	Citizen-Science	0.874	0.601	0.999
Lasius niger	Pooled	0.753	0.418	0.976
Lasius niger	Corrected C-S	0.789	0.445	0.979
Myrmica rubra	Scientific	0.924	0.735	0.948
Myrmica rubra	Citizen-Science	0.848	0.538	0.989
Myrmica rubra	Pooled	0.877	0.646	0.968
Myrmica rubra	Corrected C-S	0.916	0.704	0.971
Myrmica ruginodis	Scientific	0.805	0.437	0.981
Myrmica ruginodis	Citizen-Science	0.782	0.417	0.968
Myrmica ruginodis	Pooled	0.771	0.428	0.953
Myrmica ruginodis	Corrected C-S	0.847	0.544	0.936
Myrmica sabuleti	Scientific	0.772	0.447	0.819
Myrmica sabuleti	Citizen-Science	0.834	0.544	0.989
Myrmica sabuleti	Pooled	0.845	0.581	0.956
Myrmica sabuleti	Corrected C-S	0.937	0.769	0.96
Myrmica scabrinodis	Scientific	0.93	0.744	0.981
Myrmica scabrinodis	Citizen-Science	0.87	0.623	0.924
Myrmica scabrinodis	Pooled	0.93	0.707	0.994
Solenopsis fugax	Scientific	0.873	0.596	0.966
Solenopsis fugax	Citizen-Science	0.942	0.774	0.993
Solenopsis fugax	Pooled	0.876	0.619	0.962
Solenopsis fugax	Corrected C-S	0.954	0.802	0.979
Temnothorax nylanderi - GR	Scientific	0.948	0.771	0.96
Temnothorax nylanderi - GR	Citizen-Science	0.918	0.717	0.93
Temnothorax nylanderi - GR	Pooled	0.927	0.75	0.963
Temnothorax nylanderi - GR	Corrected C-S	0.929	0.797	0.935
Tetramorium caespitum	Scientific	0.869	0.621	0.956
Tetramorium caespitum	Citizen-Science	0.877	0.639	0.987
Tetramorium caespitum	Pooled	0.831	0.505	0.99
Tetramorium caespitum	Corrected C-S	0.863	0.562	0.981
Tetramorium impurum	Scientific	0.941	0.769	0.879
Tetramorium impurum	Citizen-Science	0.918	0.67	0.989
Tetramorium impurum	Pooled	0.924	0.712	0.94


Scientific Data
Citizen-Science Data
Difference
Pooled Datasets
F. cunicularia
F. fusca
F. lemani
F. lugubris/paralugubris
F. pressilabris
Habitat Suitability
0.00 0.25 0.50 0.75 1.00
Difference
-1.0 -0.5 0.0 0.5 1.0


Scientific Data
Citizen-Science Data
Difference
Pooled Datasets
L. flavus
L. niger
M. rubra
M. ruginodis
M. sabuleti
Habitat Suitability
0.00 0.25 0.50 0.75 1.00
Difference
-1.0 -0.5 0.0 0.5 1.0


Scientific Data
Citizen-Science Data
Difference
Pooled Datasets
M. scabrinodis
S. fugax
T. nylanderi - GR
T. caespitum
T. impurum
Habitat Suitability
0.00 0.25 0.50 0.75 1.00
Difference
-1.0 -0.5 0.0 0.5 1.0

Figures S2, S3, S4 (above). Results of ant species distribution models: i) Predicted habitat suitability using the scientific dataset, ii) predicted habitat suitability using the citizen-science dataset, iii) difference between the predictions of the citizen-science models and the scientific models, iv) predicted habitat suitability pooling the scientific dataset and a subset of the citizen-science dataset corrected for environmental bias.

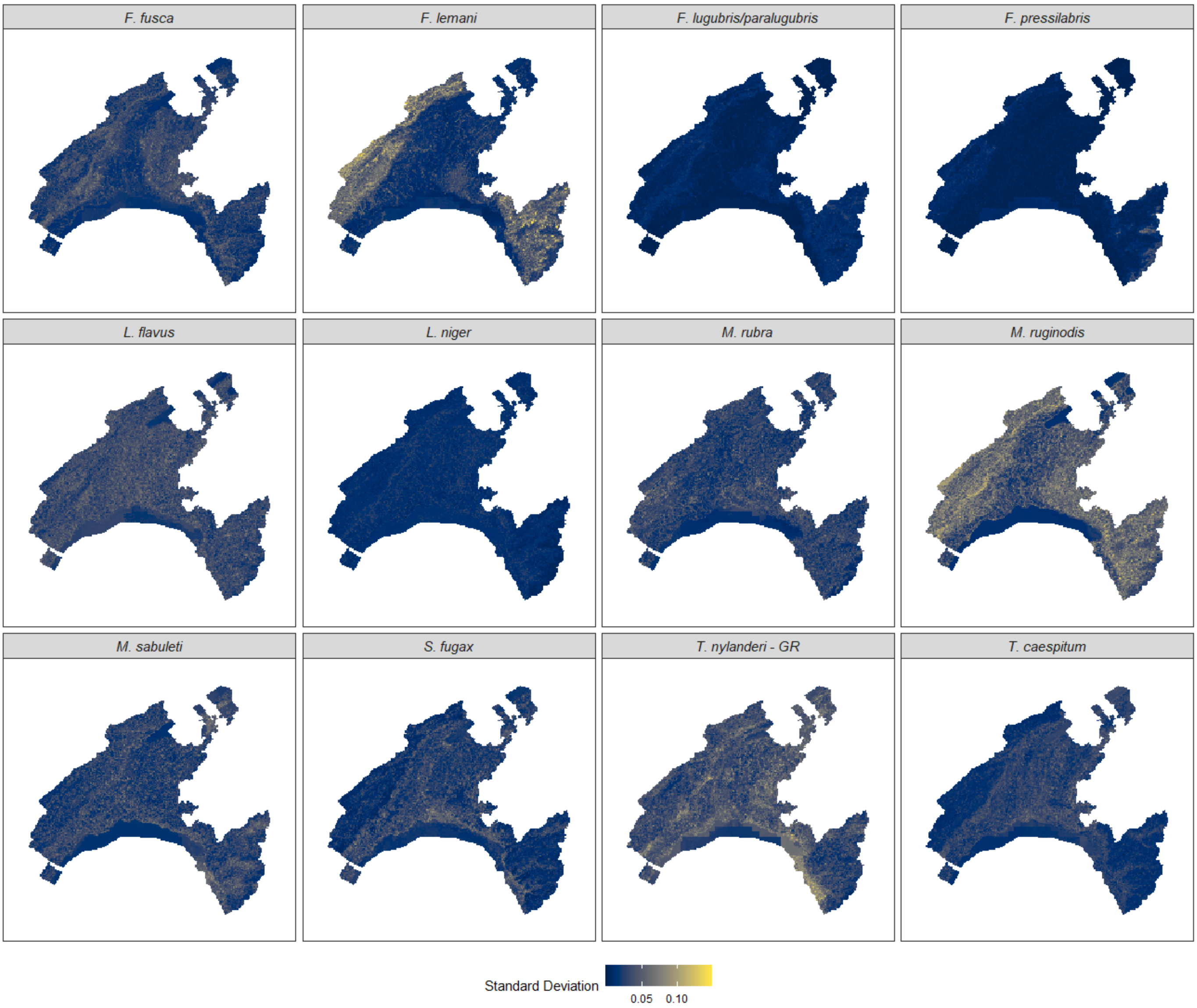


Figure S5. Variation of model predictions depending on the random choice of observations when correcting the citizen-science datasets. Standard deviation was calculated based on the results of 10 models using each a different subset of the citizen-science dataset and the same predictors.

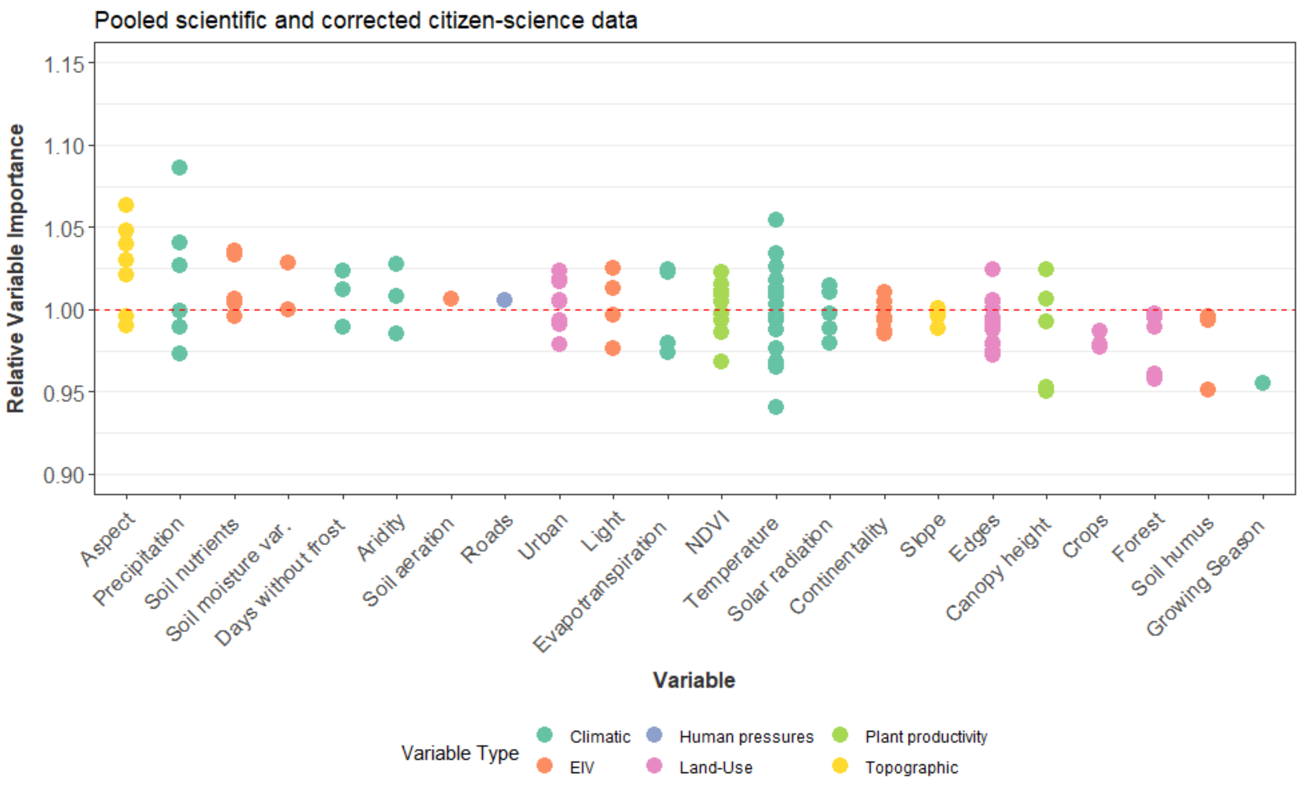


Figure S6. Relative importance of variables in the ant SDMs with the pooled scientific and corrected citizen-science dataset. Colors represent the type of environmental variable. Variables related to temperature, precipitation and habitat types were grouped to aid visualization

Regression Results

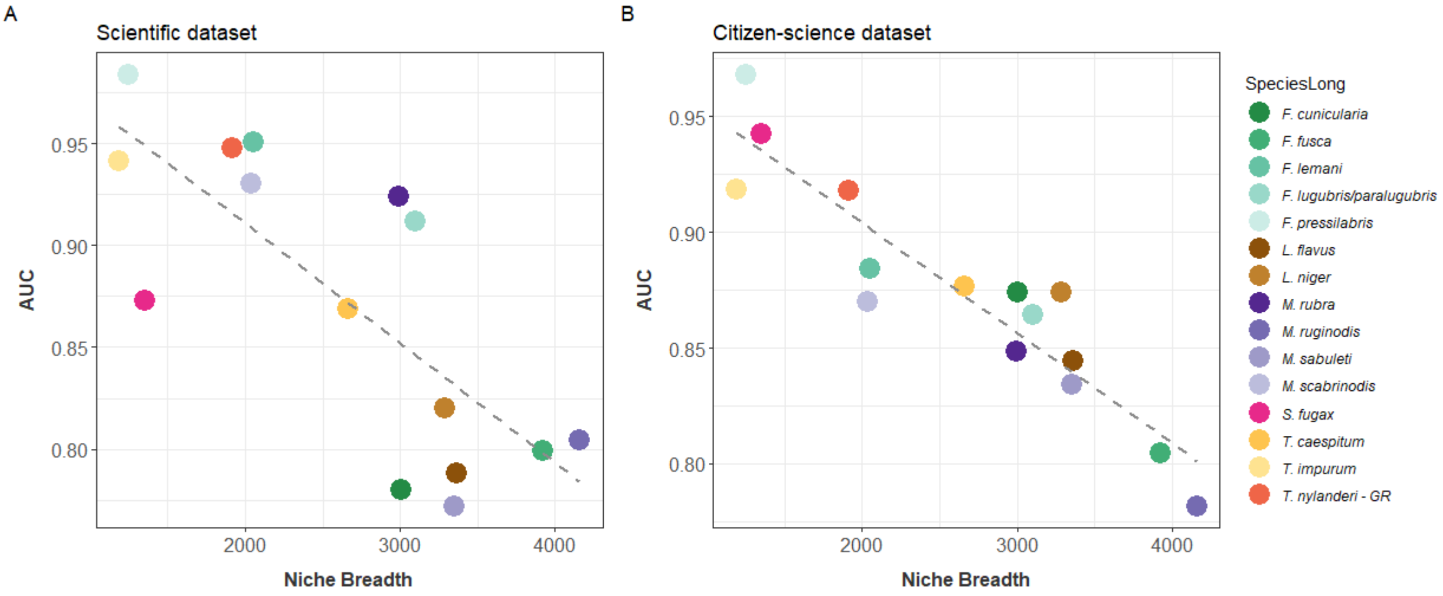


Figure S7. Simple linear regression results between niche breadth and model performance as measured with the AUC for each species. A) Performance of the models using the scientific dataset ($p < 0.05$, adjusted $R^2 = 0.5579$), B) Performance of the models using the citizen-science dataset ($p < 0.05$, adjusted $R^2 = 0.8432$).

Comparison of Spearman Correlation values

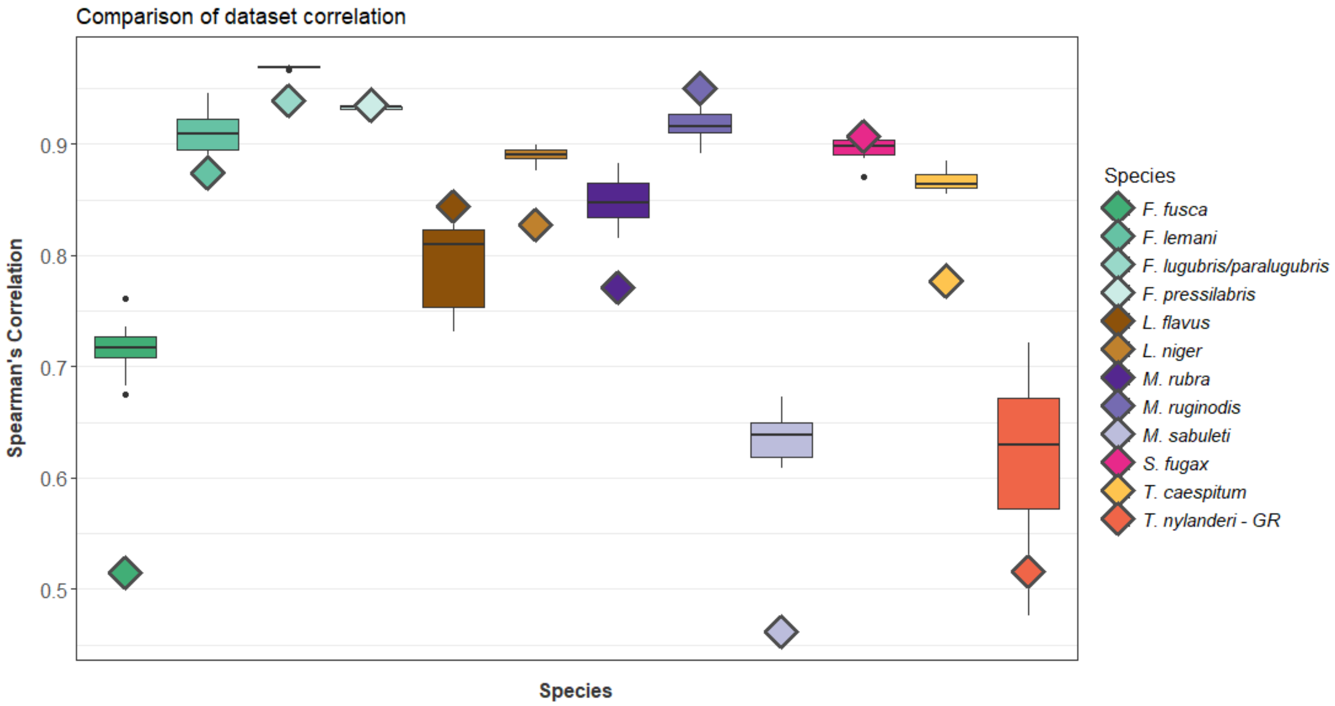


Figure S8. Comparison of spearman correlation values between habitat suitability predictions of the scientific and the citizen-science datasets i) before correction (points), ii) for 10 corrected subsets (boxplots)