Filling Monitoring Gaps for Data-deficient Species Using Annual Occupancy Predictions from Co-occurrence Models

HY Chung[a], DK Lee[b] and JE Losey[a1]

[a]*Department of Entomology at Cornell University, Ithaca, NY 14853*

[b]*Multi Campus, Seoul, South Korea 06220*

Article impact statement

Co-occurrence-based annual distribution modeling enables IUCN assessment for rare taxa, filling monitoring gaps without structured surveys.

Keywords

Data Deficient, Extinction Risk, Citizen Science, Multi-source, Data Biases, Rare Species, Insect Extinction, IUCN Red List

Word counts

5,239

Conflict of interest statement

The authors declare no conflict of interest.

Author contribution

Conceptualization Hyun Yong Chung, Dae Kyung Lee, John Losey; **Data curation** Hyun Yong Chung, Dae Kyung Lee; **Formal analysis** Hyun Yong Chung; **Funding acquisition** John Losey; **Investigation** Hyun Yong Chung, Dae Kyung Lee; **Methodology** Hyun Yong Chung, Dae Kyung Lee; **Project administration** Hyun Yong Chung; **Resources** Hyun Yong Chung, John Losey; **Software** Dae Kyung Lee, Hyun Yong Chung; **Supervision** Hyun Yong Chung; **Validation** Hyun Yong Chung; **Visualization** Hyun Yong Chung; **Writing – original draft** Hyun Yong Chung; **Writing – review and editing** Hyun Yong Chung, John Losey.

Acknowledgements

[1]**Corresponding author's email:** jel27@cornell.edu

We extend our gratitude to LM Guzman, BW Suh, and JY You for their thoughtful reviews and comments.

ORCID

Hyun Yong Chung https://orcid.org/0000-0001-7698-8105

Filling Monitoring Gaps for Data-deficient Species Using Annual Occupancy Predictions from Co-occurrence Models

Abstract

Fragmented surveys and limited monitoring have excluded most invertebrates from conservation policy. We present a fill-in framework that uses species distribution models (SDMs) to reconstruct missing annual trends—not to extrapolate trends, but to fill them in. Instead of filtering data-sparse regions or years or relying on static environmental variables, we used co-occurrence patterns (COP) as variables, to capture year-to-year assemblage shifts. COP variables enabled annual prediction at all recorded sites from multisource, presence-only, sparse data. When applied to four rare native ladybugs across North America (2007–2021), COP models exceeded reliability benchmarks (Accuracy ≥ 0.70, AUC ≥ 0.70, Kappa ≥ 0.40, Brier ≤ 0.25) across standard 7:3 splits, cross-source and cross-period validations. Annual predictions were robust to temporal biases from variation in data volume and source composition. Multiple regression indicated negligible effects of those biases on reconstructed trends. Predicted decadal declines (9–31%) closely aligned with independent regional long-term monitoring, operationalizing IUCN Red List classification (from least concern to vulnerable) in the absence of standardized monitoring. By converting fragmented observations—primarily from citizen science—into reliable annual trend estimates, the fill-in approach extends extinction-risk assessment to data-deficient taxa long excluded from conservation frameworks.

Introduction

Most invertebrate species—despite ongoing declines—remain invisible to conservation action, not because they are safe but because they are silent in the data. The International Union for Conservation of Nature (IUCN) Red List requires trend estimates over the past decade or three generations for declining species (IUCN, 2024). However, standardized monitoring programs are scarce or limited in scope for these species (Estes et al., 2018; Bayraktarov et al., 2019) and available records are typically sparse, presence-only, and opportunistic. Such data even fail to meet the requirements of current trend models (Harvey et al., 2020), which assume ordinal abundance (e.g., Newson et al., 2015; Inamine et al., 2016; Schultz et al., 2017; Martín et al., 2021), checklist-based surveys (Walker & Taylor, 2017; LeCroy et al., 2020), metrics of survey effort (e.g., Szabo et al., 2010; Isaac et al., 2014; Kamp et al., 2016; Horns et al., 2018;

63 Fink et al., 2020), and repeat visit protocols (e.g., MacKenzie et al., 2002; 2006; Kéry et al.,
64 2010; van Strien et al., 2013; Altwegg & Nichols, 2019). For example, Bayesian occupancy
65 models (OM) are effective tools for trend estimation, yet in North America they often entail ⩾
66 10,000 km² grid cells and 10–20-year intervals to achieve adequate data density for bees and
67 dragonflies (Soroye et al., 2020; Jackson et al., 2022). Without methods capable of estimating
68 temporal trends from the data available, underrepresentation of invertebrates in global
69 conservation frameworks will persist (Montgomery et al., 2020; Jönsson et al., 2021).

70 This study reconstructs annual occupancy trajectories by filling monitoring gaps with
71 annual predictions from species distribution models (SDMs) using variables derived from co-
72 occurrence patterns (COP). Previous studies have used SDMs to predict occupancy at
73 unsurveyed locations by modeling correlative relationships between known occurrence records
74 and environmental or biotic variables (Olden et al., 2008; Zimmermann et al., 2010; Franklin,
75 2013). These relationships have been modeled via recursive binary partitioning (decision-tree
76 algorithms), entropy maximization under environmental constraints (maximum entropy
77 modeling), or multivariate probit regression (joint species distribution models; JSDMs), among
78 other approaches. Such predictions can help fill monitoring gaps, and SDMs perform well with
79 small sample sizes (Hernandez et al., 2006; Wisz et al., 2008; Luan et al. 2020) and presence-
80 only data (Hernandez et al., 2006; Wisz et al., 2008; Mi et al., 2017; Robinson et al., 2018).
81 However, most applications target single-time distributions or long-term range shifts (Tingley
82 & Beissinger, 2009; Svancara et al., 2019). They primarily rely on static variables, such as
83 environmental and geographical proxies, making them ill-suited to trend estimation. Rare
84 attempts at fine-temporal prediction require structured data and extensive data thinning
85 (Svancara et al., 2019; Fink et al., 2020), which is impractical for under-monitored
86 invertebrates.

87 To predict annual occupancy at historical sites, we novelly operationalized co-occurrence
88 patterns (COP) as numerical variables for machine learning classifiers that use decision trees
89 (COP-ML). In community ecology, COP denotes the observed frequency with which species
90 co-occur across locations. Prior SDM studies suggest that such patterns can encode biotic
91 interactions (Pollock et al., 2014) and community-level environmental responses (Kissling et
92 al., 2012). Extending these ideas, we propose that annually updated COP vectors could also
93 track habitat and community shifts at a site more rapidly than static or slowly varying variables.
94 Community composition may change before environmental changes are detectable and can

immediately respond to non-environmental drivers. Here, we represented COP as a site–year vector of the annual frequencies of non-target species within a fixed radius. Decision-tree classifiers then learned rule-based partitions linking target species occupancy to both individual variables and their joint patterns (e.g., "if variable A ≥ a and variable B ≤ b, then …").

This study evaluates whether COP-ML can produce accurate predictions of annual occupancy. It further tests whether the model remains robust to temporal and structural biases in multisource, sparse and opportunistic data typical of under-monitored species. With the expansion of citizen science, opportunistic observations increasingly dominate in those taxa (Kissling et al., 2018; Knape et al., 2022) but also introduce temporal and structural biases (Isaac et al., 2014; Guzman et al., 2021; Larsen & Shirey, 2021). Mitigation strategies, such as data thinning or quality-based filtering (Wisz et al., 2008; Isaac et al., 2014; Kamp et al., 2016; Zizka et al., 2021; Van Eupen et al., 2021), can exacerbate sparsity and also preclude estimation of the absolute trend across the entire range as required under the IUCN Red List framework. Reliable annual trend estimation therefore requires methods that tolerate biased data while preserving temporal signals.

To test robustness to temporal bias, we examine whether models trained on one period generalize to others (Martínez-Minaya et al., 2018), a necessity given that citizen-science data are disproportionately concentrated in recent years (Geldmann et al., 2016). To test robustness to structural bias, we assess whether models trained on one survey type can predict others. Multisource integration is often indispensable for data-poor species (Fletcher et al., 2019; Miller et al., 2019; Isaac et al., 2020), yet inconsistent methods, even among citizen-science platforms (Gardiner et al., 2012), can introduce systematic errors (Cheney et al., 2013). For example, citizen science tends to target urban areas, whereas ecologists' surveys more often cover natural or semi-natural areas (Geldmann et al., 2016). Without such robustness, temporal trends risk reflecting shifts in prevailing survey types rather than biological change (Pagel et al., 2014; Knape et al., 2022).

We test three hypotheses: first, can COP-ML distinguish target species' presence from absence using annual COP variables? Second, can annual COP variables generalize across time periods? Third, can these variables generalize across survey methods? Accordingly, we ran three tests: standard 7:3 split evaluation, temporal generalization, and structural generalization. We then applied COP-ML to four native North American ladybugs and generated annual

occupancy estimates for 2007–2021 at all recorded sites. From these, we reconstructed trajectories and quantified 10-year reductions to operationalize IUCN Red List categories.

Methods

We asked whether COP (co-occurrence pattern) variables can yield accurate annual occupancy despite temporal and structural biases. In the step for data assembly and labeling, we compiled presence-only records from multiple sources (2007–2021) for all North American ladybugs with minimal filtering, labeling four native species as presences and two invasive competitors as pseudo-absences. In the step for variable construction, we built COP variables for each site–year labeled as presence or pseudo-absence by counting non-target species within a fixed radius and scaling counts. In the step for model training, we trained XGBoost decision-tree classifiers on these variables, which served as predictors in the model. In the step for generalization and evaluation, we assessed model performance with six metrics under three settings: cross-source (structural) generalization, cross-period (temporal) generalization, and standard 7:3 split evaluation. For each test, we summarized the distribution of performance from 2,500 resampled datasets. Finally, in the annual prediction and trend analysis step, COP-ML trained on the full dataset predicted annual occupancy of the four target species; an ensemble of resampled models (majority vote) produced the final yearly presence-absence at all historical sites for 2007–2021. From these predictions, we estimated decadal reduction rates under IUCN Criterion A and tested their significance.

Target species

Four native ladybug species—*Coccinella novemnotata* Herbst, 1793; *Coccinella transversoguttata* Faldermann, 1835; *Adalia bipunctata* (Linnaeus, 1758); and *Hippodamia parenthesis* (Say, 1824)—once dominated North American ladybug communities, thriving across diverse habitats and prey types (Losey 2007; 2012; taxonomy hereafter follows Gordon (1985) and Gordon & Vandenberg (1991)). Since the mid-1980s, their relative abundance in collections has dropped to 1–5% of former levels (Harmon et al., 2007) due to the newly-established competitors *Coccinella septempunctata* Linnaeus, 1758 and *Harmonia axyridis* (Pallas, 1773) (Wheeler & Hoebeke 1995; Harmon et al., 2007). These introduced species now dominate and alter community structure, reducing diversity and abundance continent-wide (Petersen & Losey, 2024). Estimating reduction rates and extinction risks for the natives is challenging given their current low densities and broad distributions (Wheeler & Hoebeke,

1995; Hesler et al., 2004; Harmon et al., 2007). Addressing this requires integrating multisource data across periods, regions, and survey methods while addressing inherent biases.

Occurrence data

We compiled Ladybug records from multiple sources: three citizen-science sources (The lost ladybug project, 2021; iNaturalist, 2021; BugGuide.Net, 2021), one museum website (NCSU, 2021), and three metadata platforms (GBIF, 2021; BISON, 2021; IdigBio, 2021; Appendix S1). Of the citizen-science, iNaturalist and BugGuide.Net relied on user identifications, whereas The Lost Ladybug Project relied on experts. We further verified identifications of the target species from iNaturalist and BugGuide.Net.

To assess how COP variables address biases, we applied minimal preprocessing: Records were restricted to the U.S. (excluding Alaska and Hawaii) and parts of Canada (Manitoba, Ontario, Saskatchewan, British Columbia, Alberta, Quebec) for 2007–2021. Only adult forms identified to species level were retained. Where available (89% of records), we restricted GPS accuracy to 1 km. We removed duplicate records identical in species, year, and GPS. We then computed descriptive statistics to reveal temporal and structural inconsistencies in the compiled dataset.

The dataset included 188,644 records of 353 ladybug species from 85 sources, with 324 records for *C. novemnotata*, 510 for *C. transversoguttata*, 732 for *H. parenthesis*, and 1,426 for *A. bipunctata*, which were labeled as presence.

Pseudo-absence

When explicit absence records are unavailable, pseudo-absences are often drawn by randomly sampling coordinates from all other species' records in a dataset (Robinson et al., 2018). Here we instead used records of the introduced competitors, *C. septempunctata* and *H. axyridis*. First, they competitively exclude the target species. Their occurrence within an 18-km radius without the targets was assumed to be a logical proxy for absence, reflecting a reshaped COP after local displacement. Second, their dominance (61% of our dataset) meant that traditional random sampling would largely select them anyway, minimizing methodological deviation. Therefore, we used competitors as pseudo-absences to encode a local-displacement hypothesis in COP variables.

We pooled 10,000 pseudo-absence points by subsampling from states or provinces in proportion to the regional frequency of the four targets' presence. Without this proportionality

filter, apparent accuracy increased but models relied more on geographical variables (e.g., *Coleomegilla maculata* (De Geer, 1775) concentrated in eastern regions), which are insensitive to temporal changes. In the variable construction step, we combined the entire pool with presence records. In the training and testing step, absences were resampled from this pool to balance classes.

Variables

Direct and indirect competitions shape ladybug assemblages, where the dominance of newly-established species drives niche differentiation (Petersen & Losey, 2024) and avoidance behaviors in native species (Elliott et al., 1996; Hesler & Kieckhefer 2008; Mukwevho et al., 2017). These changes can be immediate since ladybugs, as highly mobile predators, engage in long-distance interactions (e.g., *H. axyridis*, 442 km/year; McCorquodale, 1998) and actively forage across habitats (Woltz & Landis, 2013). 18 km is the commonly reported dispersal distance of this group (Jeffries et al., 2013; COSEWIC, 2016a; 2016b). Here, we represented COP as counts $c_{s,t,j}$ of cooccurring non-target ladybug species j within an 18 km radius r of site s in year t, where each site-year (s,t) was labeled as presence (= 1) or absence (= 0) of the target species. This is expressed as Equation (1):

$$c_{s,t,j} = \sum_{u} 1\,, \quad \text{if} \begin{cases} \text{species}(u) = j \\ \text{year}(u) = t \\ \text{distance}(s,u) \le r \end{cases} \tag{1}$$

Where u denotes a single georeferenced record (coordinate, year, species). Next, for each species-year (j,t), we trimmed outliers in the distribution of $c_{s,t,j}$ outside [Q1 - 1.5IQR, Q3 + 1.5 IQR] to reduce distributional bias during scaling. To maintain consistency of variable vectors, outlier detection in a single variable resulted in exclusion of the entire site–year record. This mainly affected the pseudo-absence pool, removing only 0–12 presence records per target species (≤ 0.008%). We then min-max scaled $c_{s,t,j}$ per (j,t) to adjust for species-specific overreporting and temporal variations in observation efforts, as expressed as Equation (2):

$$x_{s,t,j} = \begin{cases} \dfrac{c_{s,t,j} - \min_{s \in S_{t,j}} c_{s,t,j}}{\max_{s \in S_{t,j}} c_{s,t,j} - \min_{s \in S_{t,j}} c_{s,t,j}}, & \text{if } \max_{s \in S_{t,j}} c_{s,t,j} > \min_{s \in S_{t,j}} c_{s,t,j}\,, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We excluded environmental variables to avoid multicollinearity with COP variables, as their effects are expected to be partially embedded in co-occurrence patterns (Kissling et al., 2012). This also aligns with our focus on short-term temporal interpolation. We retained 85

species with at least 30 co-occurrences with a target species, excluding unidentified 'sp.' We applied additional screening using forward regressions ($p < 0.05$) and variance inflation factors (< 10). Although prior feature selection rarely improves performance of decision trees, we applied prescreening to limit ecologically implausible variables and to improve interpretability. To prevent leakage, we excluded target species from their own variables. Finally, we ranked the top 15 key variables using SHapley Additive exPlanations (SHAP) values, which assess feature importance in predictions from preliminary loops. The selected predictors $x_{s,t}$ were stacked into the matrix X and fed into models as:

$$X_{(s,t)} = \left(x_{s,t,j_1}, \ldots, x_{s,t,j_{15}} \right), \qquad y_{\{(s,t)\}} \in \{0,1\} \tag{3}$$

To analyze associations between each variable and target species occupancy, we calculated averages of point-biserial correlations by resampling pseudo-absence points 50 times to match presence record counts.

Development and characterization of models

We modeled associations between COP variables and targets' occupancy with XGBoost as an ensemble of decision trees:

$$F(x_i) = \sum_{m=1}^{M} f_m(x_i), \qquad \hat{p}i = \frac{1}{1 + \exp(-F(x_i))} \tag{4}$$

Here, $F(x_i)$ is the raw logit score and $\hat{p}i$ is predicted probability of presence at site-year i. Each tree f_m is a set of if–then split rules to the predictor vector x_i (the COP variables); for example, $x_a <$ Threshold split$_a$ and $x_b >$ Threshold split$_b$. Thus, tree depth $\geqslant 2$ naturally encodes interactions among predictors. Training minimizes logistic loss with a tree-complexity penalty (Chen & Guestrin 2016).

We implemented the XGBoost package in Python. From the default hyperparameter settings, we only adjusted objective='binary:logistic' and n_estimators=1000, as our aim was to evaluate COP variables rather than optimize the model.

Across all analyses, we balanced presence and pseudo-absence records 1:1 by undersampling pseudo-absences. We then generated 50 independent datasets by resampling pseudo-absences and, within each, drew 50 random train–test splits, yielding 2,500 runs (50 ×

50). For each test scenario, we summarized and evaluated performance based on the mean and distribution across runs.

We assessed model performance with six metrics: Accuracy (correct response rate), Kappa (agreement adjusted for random chance; Cohen, 1960), Recall (true positive rate), and Precision (positive predictive rate) to measure ability to predict binary presence-absence, plus Brier score (mean squared discrepancy; Brier, 1950) and AUC (ability to rank presence over absence; Fielding & Bell, 1997) for probability quality.

Generalization

Generalization tests evaluate a model's ability to predict data distinct from training data in temporal, geographical, or source aspects (Vaughan & Ormerod, 2005), minimizing train-test autocorrelation, and demonstrate robustness when ground truth comparisons are limited (Justice et al., 1999). Our tests assessed whether COP-ML could generalize across structurally or temporally distinct data pools.

Structural Generalization: To evaluate generalizability across survey types, we trained models on presence and pseudo-absence data from citizen science (LLP, iNaturalist, BugGuide.Net) to predict institutional data from 28 institutes. We assessed differences between their COP structures using ANOSIM with Manhattan distance (Appendix S2). Presence records comprised 280 citizen-driven versus 44 institutional for *C. novemnotata*, 485 versus 25 for *C. transversoguttata*, 626 versus 116 for *H. parenthesis*, and 1,338 versus 88 for *A. bipunctata*, with institutional pseudo-absence points ranging from 416 to 510. In a separate test, we also trained models on a group dominated by open-ended citizen science and evaluated them on the program emphasizing rare species (LLP).

Temporal Generalization: For forward testing, we trained models on presence data from 2007 until the year when approximately 70% of presence was accumulated, testing on the remaining about 30%. For backward testing, we reversed this, training from 2021 backward (Appendix S2). Pseudo-absence points were selected using the same cutoff year.

Evaluation

To evaluate COP-ML's annual prediction performance, we followed standard 7:3 split test by training models on 70% of presence data and testing on the remaining 30%. Unlike the generalization tests, which restricted the scope of records, this split used the full range of

presences, providing a baseline measure of COP-ML's predictive reliability for subsequent reduction rate estimation.

Prediction on annual distributions and reduction rates

To enable consistent temporal comparisons, COP-ML predicted annual presence of target species at all historical sites in our dataset since 2007, addressing yearly observation gaps.

Prediction: We developed models as described in Development and characterization of models, but we trained them on all available presence data to improve prediction accuracy given the sparsity of records (Fielding & Bell, 1997; Rencher, 1995). A site-year was classified as occupied if a majority of the 2,500 model runs (50 pseudo-absence resamples × 50 random seeds for repeated fitting) predicted presence.

Analysis: We evaluated distributional trends under IUCN Red List Criterion A, based on changes in Area of Occupancy (AOO) and Extent of Occurrence (EOO). The AOO, calculated as the number of 4 km2 grid cells occupied by a species, reflects occupancy extent and, indirectly, population size (IUCN, 2024). The EOO, defined as the polygon enclosing all known occurrences, indicates risk dispersion across a species' range (IUCN, 2024). For Criterion A, we fitted linear regressions to predicted AOO (2007–2021) and estimated the 10-year decline (2012–2021), interpreting it as a proxy for population trends (IUCN, 2024). We tested for heteroskedasticity with Breusch–Pagan and White tests, and identified influential outliers with Cook's distance. We applied robust standard errors (HC3) to assess trend significance, followed by robust regression to estimate final AOO decline. Because robust regression may downweight abrupt changes that could represent real ecological signals, we also ran ordinary least squares (OLS) regression for comparison and to improve the reliability of trend interpretation.

Validation: To verify that AOO changes predicted by COP-ML reflect consistent temporal trends despite varying data availability, we regressed predicted AOO against time (year) while including annual volumes of each citizen-science source as covariates.

Results

Biases in multisource data

The compiled dataset from multiple sources showed both structural and temporal bias. Structural bias, arising from heterogeneous effort and methods across sources (Figure 1),

appeared as unequal Efficiency (defined as the percentage of all records represented by target species): institutional data (3.5% of the compiled dataset) recorded target species 2.79 times more often than citizen-science data (96.5%). Even among citizen-science sources, LLP (5%) had Efficiency of 6.6%, 6 times that of iNaturalist where it was 1.1% (89%). Excluding LLP, the remaining dataset averaged 1.3% Efficiency (corresponding to the "lower efficiency group" in the second structural generalization). Temporal bias reflected an exponential increase in annual observations (Figure 1), with post-2014 volume being 9.61 times higher than pre-2014.

Structural and temporal generalization

We tested COP-ML for annual predictive performance and generalizability against biases through structural (cross sources) and temporal (forward/backward) generalizations. All test results met or exceeded established reliability benchmarks from prior studies: Accuracy ≥ 0.70 (rule of thumb), AUC ≥ 0.70 (Hosmer et al., 2013), Kappa ≥ 0.40 (Landis & Koch, 1977), and Brier ≤ 0.25 (Brier, 1950; Figure 2). Models trained on the citizen-science group accurately predicted presence-absence in the institution group. Likewise, models trained on the low-efficiency group generalized to the highest efficiency source, and both forward and backward temporal generalizations satisfied these benchmarks.

In the generalization from citizen-science to institutional groups, *C. transversoguttata* model achieved the highest performance (Accuracy, AUC, Kappa, Brier = 0.87, 0.94, 0.75, 0.11), followed by *C. novemnotata* (0.81, 0.85, 0.61, 0.16), *H. parenthesis* (0.78, 0.84, 0.55, 0.17), and *A. bipunctata* (0.73, 0.84, 0.46, 0.19). Analysis of similarities (ANOSIM) indicated small dissimilarities in COP predictor structures between these groups ($R < 0.25$, $p < 0.005$; Appendix S2), with species-specific R-values of 0.22 (*C. transversoguttata*), 0.12 (*H. parenthesis*), 0.06 (*C. novemnotata*), 0.05 (*A. bipunctata*), and 0.06 (pseudo-absences). Comparable results also held between the low-efficiency group, dominated by 96.3% of open-ended citizen science (plus 3.7% institutional), and the highest-efficiency source, a single citizen-science program emphasizing rare species.

In temporal generalization, *C. transversoguttata*, *C. novemnotata*, and *H. parenthesis* maintained consistent performance scores in forward and backward tests. In contrast, *A. bipunctata* showed a 7% decrease in backward Accuracy, with Recall (true positives among actual positives) increasing 2% and Precision (true positives among predicted positives) decreasing 13%. This indicates that the model trained on recent data classified a wider set of habitat conditions as occupied than were historically. In a supplementary analysis, we shifted

the split to an earlier point (pre-2012) to further exclude recent records from the test set, expanding the training set to 87% of records. This intensified the tendency: Recall rose 11% and Precision fell 18%, while accuracy remained unchanged. This suggests that its recent co-occurrence patterns encompassed conditions not occupied in the past, unlike the other three species.

Evaluation of the developed models

We assessed COP-ML classifiers, trained on 70% of the dataset and tested on 30%, against the benchmarks. All models exceeded them; even the lowest-performing species achieved Accuracy, Precision, and Recall ≥ 0.75, AUC ≥ 0.87, Kappa ≥ 0.57, and Brier ≤ 0.15 (Figure 2). *C. transversoguttata* (with 510 presence datapoints) performed best, followed by *C. novemnotata* (324), *A. bipunctata* (1,438), and *H. parenthesis* (742).

Predicted reduction rates and conservation status

Given COP-ML's demonstrated performance in tests, we generated annual occupancy predictions for 2007–2021 at all historical sites of the target species in our dataset to fill monitoring gaps and enable consistent temporal comparisons (Figure 3; Appendix S6).

Area of occupancy (AOO) declined significantly over time for all species ($p < 0.05$ for OLS and robust SE; B: -56.2 to -13.0; R^2: 0.50 to 0.83; all 95% CIs excluded zero). Although we detected heteroskedasticity in *A. bipunctata* and influential points in *A. bipunctata* and *H. parenthesis*, slope (B), R^2 and CIs were similar across OLS, OLS with robust SE, and robust regression, indicating that regression-estimated reduction rates are reliable (Appendix S3).

Predictions indicated three species were threatened by declines in North America (Appendix S4). Occupancy measured as AOO, an indicator of occupied area and an acceptable proxy of population size (IUCN, 2024), decreased from 2007 to 2021: 1,962 km^2 for *H. parenthesis*, 584 km^2 for *A. bipunctata*, and 480 km^2 each for *C. novemnotata* and *C. transversoguttata*. Under IUCN Red List Criterion A (10-year reduction), estimated rates were 31% for *H. parenthesis* (vulnerable), 15% for *A. bipunctata* (near threatened), 15% for *C. novemnotata* (near threatened), and 9% for *C. transversoguttata* (least concern; Figure 4).

Extent of occurrence (EOO), a proxy for spatial buffering against extinction risk (IUCN, 2024), declined most in *C. transversoguttata*. Despite its least concern classification here, this contraction suggests reduced spatial resilience with ongoing decline.

Multiple linear regression confirmed that time (year) was a significant predictor of AOO declines across all species, whereas annual volumes of each citizen-science source showed no evidence of statistically or practically meaningful effect (Appendix S5).

Variable importance and correlation

The importance metrics of variables (SHAP values) and point-biserial correlations showed positive co-occurrence patterns among *C. novemnotata*, *C. transversoguttata*, and *H. parenthesis.* Their occurrences were positively correlated and mutually informative in each model (Figure 5). In contrast, the two competitors *H. axyridis* and *C. septempunctata* were negatively correlated with these natives, and they were ranked among the top variables by SHAP. *A. bipunctata* was the exception: it showed positive correlations with both newly-established species. The common native *Hippodamia convergens* Guérin-Méneville, 1842 (third most abundant in the dataset) also correlated positively with three natives and ranked highly in all models, except *H. parenthesis*.

Discussion

Rationale for estimated reduction rates

This study provides the first continent-wide estimates of decadal declines for *C. novemnotata*, *C. transversoguttata*, *A. bipunctata*, and *H. parenthesis* based on annual occupancy predictions. Earlier studies from the 1980s–1990s reported steep relative abundance declines of 95–99% (rescaled from Harmon et al., 2007). By contrast, our more moderate reduction rates from 2007–2021 likely reflect the post-establishment phase in which the new competitors had already become dominant.

Several lines of evidence support the plausibility of these more moderate rates. Historical records indicate that the most acute declines occurred shortly after the establishment of *C. septempunctata* and *H. axyridis* in North America (Colunga-Garcia & Gage, 1998; Bahlai et al., 2015). Meanwhile, subsequent regional studies suggest that declines have plateaued or transitioned into a chronic, low-intensity phase (Turnock et al., 2003; Elton, 2000; Strayer et al., 2006; Harmon et al., 2007; Hesler & Kieckhefer, 2008), with no further sharp reductions observed (Alyokhin & Sewell, 2004; Bahlai et al., 2015). Such stabilization may reflect community-level reequilibration, resistance in remnant populations, or the persistence of spatial refuges (Evans, 2000; 2004; Evans et al., 2011).

Standardized long-term monitoring in Michigan (2007–2019) corroborates this interpretation, indicating 10-year declines of 37% for *H. parenthesis* and 20% for *A. bipunctata* (KBS LTER; https://lter.kbs.msu.edu/datatables/67). These trends are obtained from linear regressions of sticky-trap captures normalized by the number of survey spots to control for effort. Although limited to a single site, these local declines align closely with our continent-wide estimates (31% and 15%, respectively), suggesting that our predictions with COP-ML provide an ecologically realistic baseline for conservation assessments.

Interpretation of COP variables

Annual prediction accuracy may imply that yearly COP variables capture time-responsive ecological signals (interaction structure, habitat turnover) that static climate or land-cover variables may miss at an annual resolution. Despite attenuation bias in our noisy, heterogeneous dataset that likely damp effect sizes, correlations between key variable species and targets generally aligned with known ecological associations and were also prioritized by the models (Figure 5).

For instance, *C. novemnotata* and *C. transversoguttata* showed among the strongest positive correlations, consistent with overlapping habitat use and resource preferences (Hesler et al., 2009). In contrast, the competitors *H. axyridis* and *C. septempunctata* were negatively correlated with the three target natives and ranked among the top SHAP variables, consistent with well-documented competitive displacement (Wheeler & Hoebke, 1995; Harmon et al., 2007; Petersen & Losey, 2024).

By contrast, *A. bipunctata* showed positive correlations with both newly-established species, likely reflecting macro-scale overlap in arboreal habitat use with *H. axyridis* (Coderre et al., 1995; Koch, 2003; Omkar & Pervez, 2005; Hentley et al., 2016) and competitive coexistence with *C. septempunctata* in Europe, where both are native (Honěk 1985; Nedvěd 1999). However, this does not rule out competitive exclusion at finer spatial scales that may fall below our 18 km COP radius (Kajita et al., 2000; Kajita et al., 2006; Soares & Serpa, 2007). Notably, coexistence signals persisted despite competitor-based pseudo-absences. This alleviates concerns that pseudo-absence sampling introduced shortcut artifacts—for example, if neighborhoods of pseudo-absence points enriched with competitor records might automatically imply target absence.

Lower-ranked COP variables involved species pairs with little documented interaction and may act as proxies for geography or environment.

Strength and limitation of COP

The robustness of COP-ML across periods and sources suggests that COP variables could encode latent distributional constraints within noisy, opportunistic datasets—particularly when interactions strongly shape them (trophic dynamics, habitat filtering, or competition; Pollock et al., 2014). In our system, the prolonged competition between native and newly-established ladybugs has reshaped communities (Harmon et al., 2007; Petersen & Losey, 2024), and COP variables appear to capture these patterns.

For three species, performance metrics were similar in forward and backward predictions, indicating limited sensitivity to temporal fluctuations in data quantity and quality. Analysis of similarities (ANOSIM) reached consistent conclusions (Appendix S2). Consistent co-occurrence signals within noisy datasets provide a stable basis for annual occupancy prediction. One possible explanation is that COP variables capture relational signals, which are less sensitive to sampling noise than single-species occurrence rates (Tikhonov et al., 2017; Johnston et al., 2017).

However, this strength depends on the temporal stability of co-occurrence patterns (Tikhonov et al., 2017). COP-ML declined in performance when backcasting *A. bipunctata*, whose habitat selection was reported to have shifted under post-invasion habitat compression (Bahlai et al., 2015). The model trained on recent COP tended to overpredict past suitability by classifying historically unoccupied conditions as suitable—evidenced by higher recall than precision. This underscores a limitation: when biotic interactions change, the assumption of time-invariant COP may fail.

Our COP-ML generalized across heterogeneous datasets—from open-ended to targeted rare-species citizen science—while showing small dissimilarities in COP values (Appendix S2). Although opportunistic data are often viewed skeptically and has limited its utility (Isaac & Pocock, 2015; Steen et al., 2019), multisource integration is increasingly essential for data-poor species (Miller et al., 2019; Isaac et al., 2020). Our results indicate that COP variables, largely driven by commonly recorded species, can indirectly inform distributions of rarer taxa. These findings highlight the conservation value of citizen science and suggest that its rapidly growing data volumes can be productively leveraged.

Fill-in approach with annual predictions

We propose a fill-in approach that generates predictions of annual occupancy to bridge monitoring gaps. By tracking year-to-year occupancy across North America since 2007, we evaluated 10-year reductions under the IUCN Red List Criterion A (Figure 3). Traditional time-series workflows often filter datasets to well-monitored regions, narrowing the spatiotemporal scope of inference and precluding absolute-extent assessments (e.g., IUCN Red List). Current trend models require extensive filtering or structured surveys—resources structurally inaccessible to the taxa most in need of conservation insight. Our approach produces fine-temporal estimations from sparse, presence-only, multisourced datasets, directly benefitting them.

For under-monitored species, annual COP-ML predictions could complement application of Bayesian occupancy models (OM) that track temporal change but typically require at least two revisits per period or high spatiotemporal density in data (Royle, 2006; Kamp et al., 2016; Outhwaite et al., 2018; Perkins-Taylor & Frey, 2020; Jha et al., 2022). In North America, limited data density has often forced coarse spatial (~10,000 km²) and temporal (10–20 years) resolutions for insects such as bees and dragonflies (Soroye et al., 2020; Jackson et al., 2022; Shirey et al., 2023). One way to recover resolution is to commission additional, targeted surveys (Xue et al., 2016; Tulloch et al., 2013), but this is costly; our approach offers an alternative by producing annual predictions without new field effort. However, incorporating ML-based predictions into occupancy modeling frameworks—as pseudo-observations—remains largely untested. Occupancy models explicitly model detection and survey processes, so predicted probabilities must be reconciled with those components. Although recent OM advances have explored non-ideal data (e.g., assuming random-walk observation processes, using pseudo-absence instead of checklist absence, or treating opportunistic records as revisits; Outhwaite et al., 2018), the statistical compatibility of ML predictions within OM frameworks has yet to be demonstrated.

The fill-in and filtering strategies are complementary. However, we deliberately did not apply spatial thinning or filtering to address spatial autocorrelation. First, our targets are data-deficient, and our goal is to cover the entire known range, so filtering is impractical. Second, our aim was to test whether COP-ML, devised for such taxa, remains reliable without deep filtering. Third, our case is temporal interpolation at the same sites, rather than prediction to new locations, spatial leakage is less relevant in our setting. Finally, state-matched pseudo-absences further mitigate overfitting by emphasizing within-region discrimination. Future

study will be needed to determine filtering levels that optimally balance performance, bias reduction, and data retention in the fill-in framework.

This study presents a scalable method to bridge monitoring gaps for data-deficient species, using sparse, largely opportunistic, presence-only records to generate annual occupancy estimates. Predicted trends aligned with long-term trends from independent regional monitoring and operationalized IUCN Red List criteria for species previously excluded due to lack of data. COP-ML demonstrated robust performance across heterogeneous sources and time periods, showing that reliable signals of extinction risk can emerge even from unstructured datasets. By converting fragmented observations into interpretable trends, the fill-in approach with annual predictions provides a practical pathway to extend assessment of extinction risk and strengthen conservation decisions where standardized monitoring is absent.

References

Altwegg, R., & Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, **10(1)**, 8-21.

Alyokhin, A., & Sewell, G. (2004). Changes in a lady beetle community following the establishment of three alien species. *Biological Invasions*, **6(4)**, 463–471.

Bahlai, C. A., Colunga-Garcia, M., Gage, S. H., & Landis, D. A. (2015). The role of exotic ladybeetles in the decline of native ladybeetle populations: evidence from long-term monitoring. *Biological Invasions*, **17**, 1005-1024.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution*, **6**, 239.

Biodiversity Information Serving Our Nation (BISON). (2021, September 16). Coccinellidae records, 2007–2021, United States and Canada [Data set]. Retrieved from http://bison.usgs.ornl.gov/

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78(1)**, 1-3.

BugGuide.Net. (2021, September 26). Coccinellidae search results [Web-scraped data]. Retrieved from https://bugguide.net/node/view/15740

514 Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In B.
515 Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings*
516 *of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*
517 *Mining* (pp. 785–794). ACM.

518 Cheney, B., Thompson, P. M., Ingram, S. N., Hammond, P. S., Stevick, P. T., Durban, J. W.,
519 Culloch, R. M., Elwen, S. H., Mandleberg, L., Janik, V. M., Quick, N. J., Islas-Villanueva, V.,
520 Robinson, K. P., Costa, M., Eisfeld, S. M., Walters, A., Philips, C., Weir, C. R., Evans, P. G.
521 H., ... Wilson, B. (2013). Integrating multiple data sources to assess the distribution and
522 abundance of bottlenose dolphins Tursiops truncatus in Scottish waters. *Mammal Review*, **43(1)**,
523 71–88.

524 Coderre, D., Lucas, É., & Gagné, I. (1995). The occurrence of *Harmonia axyridis*
525 (Pallas)(Coleoptera: Coccinellidae) in Canada. *The Canadian Entomologist*, **127(4)**, 609-611.

526 Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological*
527 *measurement*, **20(1)**, 37-46.

528 Colunga-Garcia, M., & Gage, S. H. (1998). Arrival, establishment, and habitat use of the
529 multicolored Asian lady beetle (Coleoptera: Coccinellidae) in a Michigan landscape.
530 *Environmental Entomology*, **27(6)**, 1574-1580.

531 COSEWIC. (2016a). COSEWIC assessment and status report on the Transverse Lady Beetle
532 (*Coccinella transversoguttata*) in Canada. Committee on the Status of Endangered Wildlife in
533 Canada.

534 COSEWIC. (2016b). COSEWIC assessment and status report on the Nine-spotted Lady Beetle
535 (*Coccinella novemnotata*) in Canada. Committee on the Status of Endangered Wildlife in
536 Canada.

537 Elliott, N., Kieckhefer, R., & Kauffman, W. (1996). Effects of an invading coccinellid on native
538 coccinellids in an agricultural landscape. *Oecologia*, **105(4)**, 537–544.

539 Elton, C. S. (2000). The ecology of invasions by animals and plants. *Methuen*. (Original work
540 published 1958)

541 Estes, L., Elsen, P. R., Treuer, T., Ahmed, L., Caylor, K., Chang, J., Choi, J. J., & Ellis, E. C.
542 (2018). The spatial and temporal domains of modern ecology. *Nature Ecology & Evolution*,
543 **2(5)**, 819–826.

544 Evans, E. W. (2000). Morphology of invasion: body size patterns associated with establishment
545 of *Coccinella septempunctata* (Coleoptera: Coccinellidae) in western North America.
546 *European Journal of Entomology*, **97(4)**, 469-474.

547 Evans, E. W. (2004). Habitat displacement of North American ladybirds by an introduced
548 species. *Ecology*, **85(3)**, 637-647.

549 Evans, E. W., Soares, A. O., & Yasuda, H. (2011). Invasions by ladybugs, ladybirds, and other
550 predatory beetles. *BioControl*, **56**, 597-611.

551 Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction
552 errors in conservation presence/absence models. *Environmental conservation*, **24(1)**, 38-49.

553 Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020).
554 Modeling avian full annual cycle distribution and population trends with citizen science data.
555 *Ecological Applications*, **30(3)**, e02056.

556 Fletcher Jr, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio,
557 R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*,
558 **100(6)**, e02710.

559 Franklin, J. (2013). Species distribution models in conservation biogeography: developments
560 and challenges. *Diversity and distributions*, **19(10)**, 1217-1223.

561 Gardiner, M. M., Allee, L. L., Brown, P. M., Losey, J. E., Roy, H. E., & Smyth, R. R. (2012).
562 Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science
563 programs. *Frontiers in Ecology and the Environment*, **10(9)**, 471-476.

564 GBIF. (2022, October 31). Coccinellidae occurrence dataset, 2007–2021, United States and
565 Canada [Data set]. GBIF Occurrence Download. https://doi.org/10.15468/dl.vgen57

566 Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B. O., Olsen, K.,
567 Rahbek, C., & Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring
568 four recording schemes with different proficiency requirements. *Diversity and Distributions*,
569 **22(11)**, 1139–1149.

570 Gordon, R. D. (1985). The Coccinellidae (Coleoptera) of America North of Mexico. *Journal*
571 *of the New York Entomological Society*, **93(1)**, 1-912.

572 Gordon, R. D., & Vandenberg, N. J. (1991). Field guide to recently introduced species of
573 Coccinellidae (Coleoptera) in North America, with a revised key to North American genera of
574 Coccinellini. *Proceedings of the Entomological Society of Washington*, **93(4)**, 845-867.

575 Guzman, L. M., Johnson, S. A., Mooers, A. O., & M'Gonigle, L. K. (2021). Using historical
576 data to estimate bumble bee occurrence: Variable trends across species provide little support
577 for community-level declines. *Biological Conservation*, **257**, 109141.

578 Harmon, J. P., Stephens, E., & Losey, J. (2007). The decline of native coccinellids (Coleoptera:
579 Coccinellidae) in the United States and Canada. *Journal of Insect Conservation*, **11(2)**, 85–94.

580 Harvey, J. A., Heinen, R., Armbrecht, I., Basset, Y., Baxter-Gilbert, J. H., Bezemer, T. M.,
581 Böhm, M., Christie, A. P., Cornelisse, T., Crone, E. E., Dicke, M., Dicks, L. V., Elder, M.,
582 Fartmann, T., Forister, M. L., Gaston, K. J., Jepsen, S. J., Jones, T. H., Kaydan, M. B., ... de
583 Kroon, H. (2020). International scientists formulate a roadmap for insect conservation and
584 recovery. *Nature Ecology & Evolution*, **4(2)**, 174–176.

585 Hentley, W. T., Vanbergen, A. J., Beckerman, A. P., Brien, M. N., Hails, R. S., Jones, T. H., &
586 Johnson, S. N. (2016). Antagonistic interactions between an invasive alien and a native
587 coccinellid species may promote coexistence. *Journal of Animal Ecology*, **85(4)**, 1087-1097.

588 Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample
589 size and species characteristics on performance of different species distribution modeling
590 methods. *Ecography*, **29(5)**, 773-785.

591 Hesler, L. S., & Kieckhefer, R. W. (2008). Status of exotic and previously common native
592 coccinellids (Coleoptera) in South Dakota landscapes. *Journal of the Kansas Entomological*
593 *Society*, **81(1)**, 29-49.

594 Hesler, L. S., Catangui, M. A., Losey, J. E., Helbig, J. B., & Mesman, A. (2009). Recent records
595 of *Adalia bipunctata* (L.), *Coccinella transversoguttata* richardsoni Brown, and *Coccinella*
596 *novemnotata* Herbst (Coleoptera: Coccinellidae) from South Dakota and Nebraska. *The*
597 *Coleopterists Bulletin*, **63(4)**, 475-484.

598 Hesler, L. S., Kieckhefer, R. W., & Catangui, M. A. (2004). Surveys and field observations of
599 *Harmonia axyridis* and other Coccinellidae (Coleoptera) in eastern and central South Dakota.
600 *Transactions of the American Entomological Society*, **130(1)**, 113–133.

Honěk, A. (1985). Habitat preferences of aphidophagous coccinellids [Coleoptera]. *Entomophaga*, **30(3)**, 253-264.

Horns, J. J., Adler, F. R., & Şekercioğlu, Ç. H. (2018). Using opportunistic citizen science data to estimate avian population trends. *Biological conservation*, **221**, 151-159.

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). *Wiley.*

iDigBio. (2021, September 16). Coccinellidae records, 2007–2021, United States and Canada [Data set]. Retrieved from https://www.idigbio.org/

Inamine, H., Ellner, S. P., Springer, J. P., & Agrawal, A. A. (2016). Linking the continental migratory cycle of the monarch butterfly to understand its population decline. *Oikos*, **125(8)**, 1081-1091.

iNaturalist. (2021, September 28). Coccinellidae observations, 2007–2021, United States and Canada [Data set]. Retrieved from https://www.inaturalist.org/

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.

Isaac, N. J., & Pocock, M. J. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, **115(3)**, 522-531.

Isaac, N. J., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, **5(10)**, 1052-1060.

IUCN. (2024). The IUCN Red List of Threatened Species. Version 16. Available at: https://www.iucnredlist.org/resources/redlistguidelines.

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, **9(1)**, 88-97.

629 Jackson, H. M., Johnson, S. A., Morandin, L. A., Richardson, L. L., Guzman, L. M., &
630 M'Gonigle, L. K. (2022). Climate change winners and losers among North American
631 bumblebees. *Biology letters*, **18(6)**, 20210551.

632 Jeffries, D. L., Chapman, J., Roy, H. E., Humphries, S., Harrington, R., Brown, P. M., &
633 Handley, L. J. L. (2013). Characteristics and drivers of high-altitude ladybird flight: insights
634 from vertical-looking entomological radar. *PloS one*, **8(12)**, e82278.

635 Jha, A., Praveen, J., & Nameer, P. O. (2022). Contrasting occupancy models with presence-
636 only models: does accounting for detection lead to better predictions?. *Ecological Modelling*,
637 **472**, 110105.

638 Jönsson, G. M., Broad, G. R., Sumner, S., & Isaac, N. J. (2021). A century of social wasp
639 occupancy trends from natural history collections: spatiotemporal resolutions have little effect
640 on model performance. *Insect Conservation and Diversity*, **14(5)**, 543-555.

641 Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of
642 prognostic information. *Annals of internal medicine*, **130(6)**, 515-524.

643 Kajita, Y., Takano, F., Yasuda, H., & Agarwala, B. K. (2000). Effects of indigenous ladybird
644 species (Coleoptera: Coccinellidae) on the survival of an exotic species in relation to prey
645 abundance. *Applied Entomology and Zoology*, **35(4)**, 473-479.

646 Kajita, Y., Yasuda, H., & Evans, E. W. (2006). Effects of native ladybirds on oviposition of the
647 exotic species, *Adalia bipunctata* (Coleoptera: Coccinellidae), in Japan. *Applied Entomology*
648 *and Zoology*, **41(1)**, 57-61.

649 Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., & Donald, P. F. (2016). Unstructured citizen
650 science data fail to detect long-term population declines of common birds in Denmark.
651 *Diversity and Distributions*, **22(10)**, 1024-1035.

652 Kéry, M., Gardner, B., & Monnerat, C. (2010). Predicting species distributions from checklist
653 data using site-occupancy models. *Journal of Biogeography*, **37(10)**, 1851-1862.

654 Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernández, N., García, E. A.,
655 Guralnick, R. P., Isaac, N. J. B., Kelling, S., Los, W., McRae, L., Mihoub, J. B., Obst, M.,
656 Santamaria, M., Skidmore, A. K., Williams, K. J., Agosti, D., Amariles, D., Arvanitidis, C.,
657 Bastin, L., ... Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species
658 distribution and abundance at a global scale. *Biological Reviews*, **93**, 600–625.

659 Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G. J., Montoya,
660 J. M., Römermann, C., Schiffers, K., Schurr, F. M., Singer, A., Svenning, J.-C., Zimmermann,
661 N. E., & O'Hara, R. B. (2012). Towards novel approaches to modelling biotic interactions in
662 multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39(12)**, 2163–2178.

663 Knape, J., Coulson, S. J., van der Wal, R., & Arlt, D. (2022). Temporal trends in opportunistic
664 citizen science reports across multiple taxa. *Ambio*, **51(1)**, 183–198.

665 Koch, R. L. (2003). The multicolored Asian lady beetle, *Harmonia axyridis*: a review of its
666 biology, uses in biological control, and non-target impacts. *Journal of insect Science*, **3(1)**, 32.

667 Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical
668 data. *Biometrics*, **33(1)**, 159–174.

669 Larsen, E. A., & Shirey, V. (2021). Method matters: Pitfalls in analysing phenology from
670 occurrence records. *Ecology Letters*, **24(6)**, 1287-1289.

671 LeCroy, K. A., Savoy-Burke, G., Carr, D. E., Delaney, D. A., & Roulston, T. A. H. (2020).
672 Decline of six native mason bee species following the arrival of an exotic congener. *Scientific*
673 *reports*, **10(1)**, 18745.

674 Losey, J. E., Perlman, J. E., & Hoebeke, E. R. (2007). Citizen scientist rediscovers rare nine-
675 spotted lady beetle, *Coccinella novemnotata*, in eastern North America. *Journal of Insect*
676 *Conservation*, **11(4)**, 415–417.

677 Losey, J., Allee, L., & Smyth, R. (2012). The Lost Ladybug Project: Citizen spotting surpasses
678 scientist's surveys. *American Entomologist*, **58(1)**, 22-24.

679 Luan, J., Zhang, C., Xu, B., Xue, Y., & Ren, Y. (2020). The predictive performances of random
680 forest models with limited sample size and different species traits. *Fisheries Research*, **227**,
681 105534.

682 MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm,
683 C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one.
684 *Ecology*, **83(8)**, 2248-2255.

685 MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., & Hines, J. E. (2017).
686 Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence
687 (2nd ed.). *Academic Press*.

688 Martín, B., González–Arias, J., & Vicente–Vírseda, J. A. (2021). Machine learning as a
689 successful approach for predicting complex spatio–temporal patterns in animal species
690 abundance. *Machine learning*, **44**, 289-301.

691 Martínez-Minaya, J., Cameletti, M., Conesa, D., & Pennino, M. G. (2018). Species distribution
692 modeling: a statistical review with focus in spatio-temporal issues. *Stochastic environmental*
693 *research and risk assessment*, **32**, 3227-3244.

694 McCorquodale, D. B. (1998). Adventive lady beetles (Coleoptera: Coccinellidae) in eastern
695 Nova Scotia, Canada. *Entomological News*, **109**, 15–20.

696 Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to
697 predict rare species distribution with few samples in large undersampled areas? Three Asian
698 crane species models provide supporting evidence. *PeerJ*, **5**, e2849.

699 Miller, D. A., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising
700 future for data integration methods to estimate species' distributions. *Methods in Ecology and*
701 *Evolution*, **10(1)**, 22-37.

702 Montgomery, G. A., Dunn, R. R., Fox, R., Jongejans, E., Leather, S. R., Saunders, M. E.,
703 Shortall, C. R., Tingley, M. W., & Wagner, D. L. (2020). Is the insect apocalypse upon us? How
704 to find out. *Biological Conservation*, **241**, 108327.

705 Mukwevho, V. O., Pryke, J. S., & Roets, F. (2017). Habitat preferences of the invasive
706 harlequin ladybeetle *Harmonia axyridis* (Coleoptera: Coccinellidae) in the Western Cape
707 Province, South Africa. *African Entomology*, **25(1)**, 86-97.

708 Nedvěd O. 1999. Host complexes of predaceous ladybeetles (Coleoptera: Coccinellidae).
709 *Journal of Applied Entomology*, **123**, 73–76.

710 Newson, S. E., Evans, H. E., & Gillings, S. (2015). A novel citizen science approach for large-
711 scale standardised monitoring of bat activity and distribution, evaluated in eastern England.
712 *Biological Conservation*, **191**, 38-49.

713 North Carolina State University Insect collection (NCSU). (2021, September 26).
714 Coccinellidae records [Web-scraped data]. Retrieved from
715 http://specimens.insectmuseum.org/public/specimen

716 Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine learning methods without tears: a
717 primer for ecologists. *The Quarterly review of biology*, **83(2)**, 171-193.

718 Omkar, & Pervez, A. (2005). Ecology of two-spotted ladybird, *Adalia bipunctata*: a review.
719 *Journal of Applied Entomology*, **129(9-10)**, 465-474.

720 Outhwaite, C. L., Chandler, R. E., Powney, G. D., Collen, B., Gregory, R. D., & Isaac, N. J.
721 (2018). Prior specification in Bayesian occupancy modelling improves analysis of species
722 occurrence data. *Ecological Indicators*, **93**, 333-343.

723 Pagel, J., Anderson, B. J., O'Hara, R. B., Cramer, W., Fox, R., Jeltsch, F., Roy, D. B., Thomas,
724 C. D., & Schurr, F. M. (2014). Quantifying range-wide variation in population trends from local
725 abundance surveys and widespread opportunistic occurrence records. *Methods in Ecology and*
726 *Evolution*, **5(8)**, 751–760.

727 Perkins-Taylor, I. E., & Frey, J. K. (2020). Predicting the distribution of a rare chipmunk
728 (Neotamias quadrivittatus oscuraensis): comparing MaxEnt and occupancy models. *Journal of*
729 *Mammalogy*, **101(4)**, 1035-1048.

730 Petersen, M. J., & Losey, J. E. (2024). Niche overlap with an exotic competitor mediates the
731 abundant niche-centre relationship for a native lady beetle. *Diversity and Distributions*, **30(5)**,
732 e13825.

733 Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A.,
734 & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously
735 with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5(5)**,
736 397–406.

737 Rencher, A. C. (1995). Methods of multivariate analysis*. Wiley.*

738 Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution
739 modelling for rare species using citizen science data. *Diversity and Distributions*, **24(4)**, 460-
740 472.

741 Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities.
742 *Biometrics*, **62(1)**, 97-102.

743 Schultz, C. B., Brown, L. M., Pelton, E., & Crone, E. E. (2017). Citizen science monitoring
744 demonstrates dramatic declines of monarch butterflies in western North America. *Biological*
745 *Conservation*, **214**, 343-346.

Shirey, V., Khelifa, R., M'Gonigle, L. K., & Guzman, L. M. (2023). Occupancy–detection models with museum specimen data: Promise and pitfalls. *Methods in Ecology and Evolution*, **14(2)**, 402-414.

Soares, A. O., & Serpa, A. (2007). Interference competition between ladybird beetle adults (Coleoptera: Coccinellidae): Effects on growth and reproductive capacity. *Population Ecology*, **49(1)**, 37–43.

Soroye, P., Newbold, T., & Kerr, J. (2020). Climate change contributes to widespread declines among bumble bees across continents. *Science*, **367(6478)**, 685-688.

Steen, V. A., Elphick, C. S., & Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions*, **25(12)**, 1857-1869.

Strayer, D. L., Eviner, V. T., Jeschke, J. M., & Pace, M. L. (2006). Understanding the long-term effects of species invasions. *Trends in ecology & evolution*, **21(11)**, 645-651.

Svancara, L. K., Abatzoglou, J. T., & Waterbury, B. (2019). Modeling current and future potential distributions of milkweeds and the monarch butterfly in Idaho. *Frontiers in Ecology and Evolution*, **7**, 168.

Szabo, J. K., Vesk, P. A., Baxter, P. W., & Possingham, H. P. (2010). Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. *Ecological Applications*, **20(8)**, 2157-2169.

The Lost Ladybug Project. (2021, September 28). Lost Ladybug Project dataset [Data set]. Retrieved from http://www.lostladybug.org/

Tikhonov, G., Abrego, N., Dunson, D., & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, **8(4)**, 443-452.

Tingley, M. W., & Beissinger, S. R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in ecology & evolution*, **24(11)**, 625-633.

Tulloch, A. I., 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.

775 Turnock, W. J., Wise, I. L., & Matheson, F. O. (2003). Abundance of some native coccinellines
776 (Coleoptera: Coccinellidae) before and after the appearance of *Coccinella septempunctata*. *The*
777 *Canadian Entomologist*, **135(3)**, 391-404.

778 Van Eupen, C., Maes, D., Herremans, M., Swinnen, K. R., Somers, B., & Luca, S. (2021). The
779 impact of data quality filtering of opportunistic citizen science data on species distribution
780 model performance. *Ecological Modelling*, **444**, 109453.

781 Van Strien, A. J., Van Swaay, C. A., & Termaat, T. (2013). Opportunistic citizen science data
782 of animal species produce reliable estimates of distribution trends if analysed with occupancy
783 models. *Journal of Applied Ecology*, **50(6)**, 1450-1458.

784 Vaughan, I. P., & Ormerod, S. J. (2005). The continuing challenges of testing species
785 distribution models. *Journal of applied ecology*, **42(4)**, 720-730.

786 Walker, J., & Taylor, P. D. (2017). Using eBird data to model population change of migratory
787 bird species. *Avian Conservation and Ecology*, **12(1)**, 4.

788 Wheeler, A. G., Jr., & Hoebeke, E. R. (1995). *Coccinella novemnotata* in northeastern North
789 America: Historical occurrence and current status (Coleoptera: Coccinellidae). *Proceedings of*
790 *the Entomological Society of Washington*, **97**, 701–716.

791 Willig, M. R., Woolbright, L., Presley, S. J., Schowalter, T. D., Waide, R. B., Heartsill Scalley,
792 T., Zimmerman, J. K., González, G., & Lugo, A. E. (2019). Populations are not declining and
793 food webs are not collapsing at the Luquillo Experimental Forest. *Proceedings of the National*
794 *Academy of Sciences*, **116(25)**, 12143–12144.

795 Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS
796 Predicting Species Distributions Working Group. (2008). Effects of sample size on the
797 performance of species distribution models. *Diversity and distributions*, **14(5)**, 763-773.

798 Woltz, J. M., & Landis, D. A. (2013). Coccinellid immigration to infested host patches
799 influences suppression of Aphis glycines in soybean. *Biological Control*, **64(3)**, 330-337.

800 Xue, Y., Davies, I., Fink, D., Wood, C., & Gomes, C. P. 2016. Avicaching: A two stage game
801 for bias reduction in citizen science. *Proceedings of the 15th International Conference on*
802 *Autonomous Agents and Multiagent Systems*, 776–785.

803 Zimmermann, N. E., Edwards Jr, T. C., Graham, C. H., Pearman, P. B., & Svenning, J. C.
804 (2010). New trends in species distribution modelling. *Ecography*, **33(6)**, 985-989.

805 Zizka, A., Antonelli, A., & Silvestro, D. (2021). Sampbias, a method for quantifying geographic
806 sampling biases in species distribution data. *Ecography*, **44(1)**, 25-32.

807

Figures

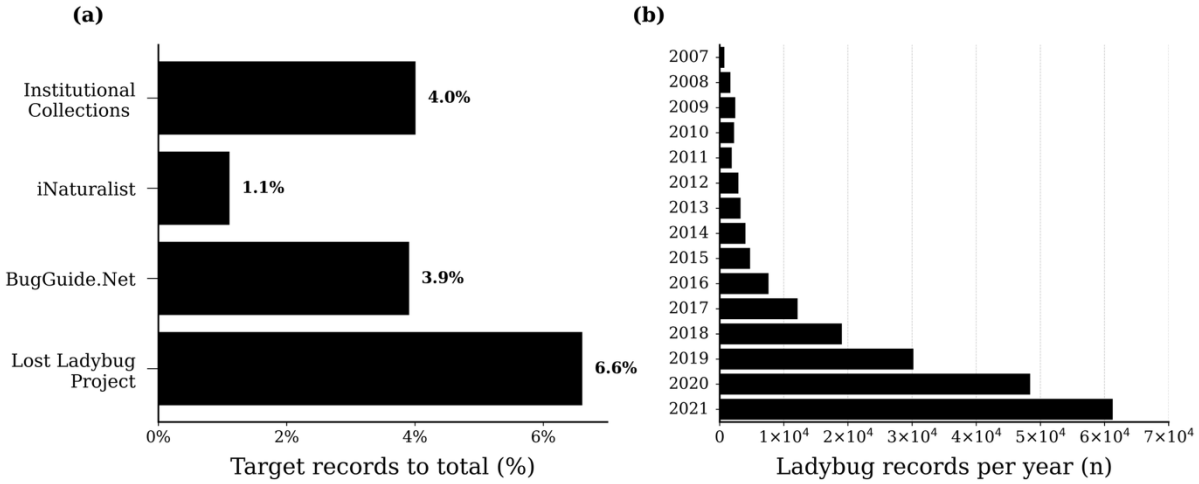
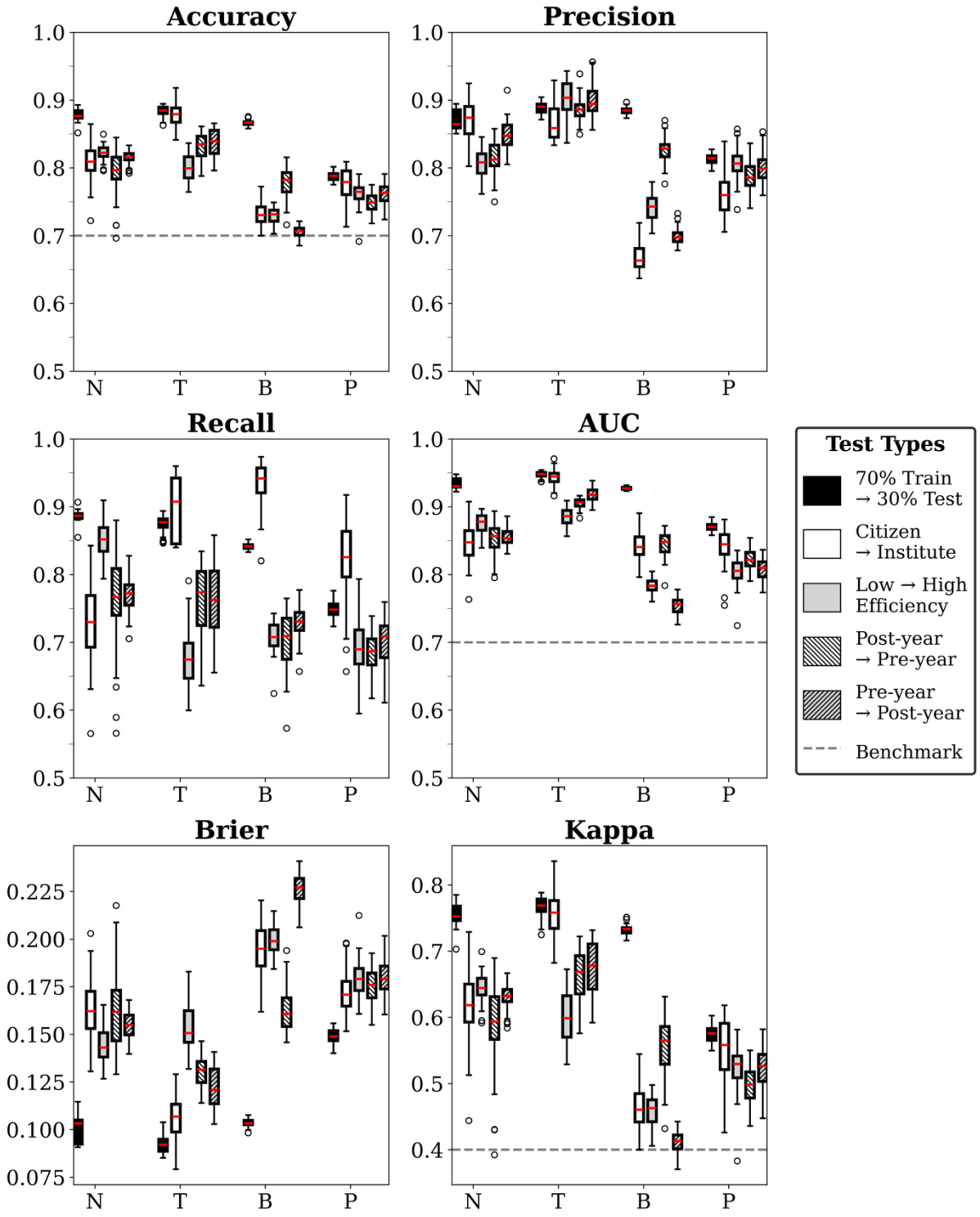


Figure 1. Structural and temporal inconsistencies in our multisourced dataset; **(a)** difference in detection efficiency (target records per 100 ladybug records) in institutional collections and three citizen science sources (The Lost Ladybug Project, iNaturalist, BugGuide.Net); **(b)** exponential increase in annual ladybug observations across the USA and Canada (2007–2021; $y = 758.89e^{0.29x}$, $R^2 = 0.88$).



817

818 **Figure 2.** Performance of species distribution models using co-occurrence pattern variables for
819 *C. novemnotata* (N), *C. transversoguttata* (T), *A. bipunctata* (B), and *H. parenthesis* (P) across
820 standard evaluation, structural generalizations, and temporal generalizations; black plots,
821 standard 70% training and 30% testing; white plots, citizen science training and institutional
822 data testing; gray plots, low target-density group (1.3%) training and high density source (6.6%)
823 testing; left hatch, post-2007 training until ~70% coverage and later-year testing; right hatch,

Pre-2021 traing until ~70% coverage and earlier-year testing; red lines, mean performance across 2,500 runs; dash, reliability benchmarks—Accuracy ≥ 0.70, AUC ≥ 0.70 (Hosmer et al., 2013), Kappa ≥ 0.40 (Landis & Koch, 1977), and Brier ≤ 0.25 (Brier, 1950).

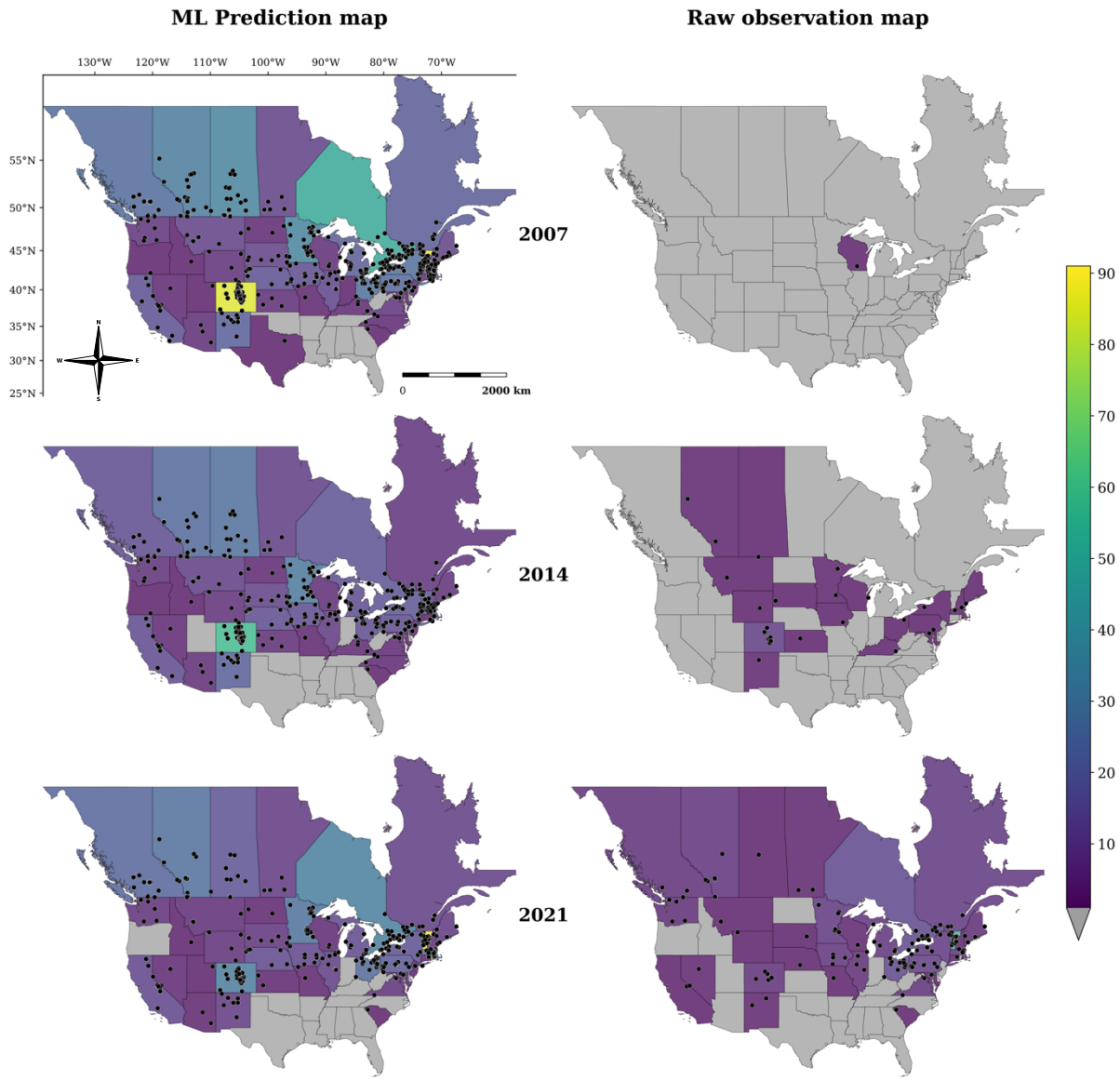
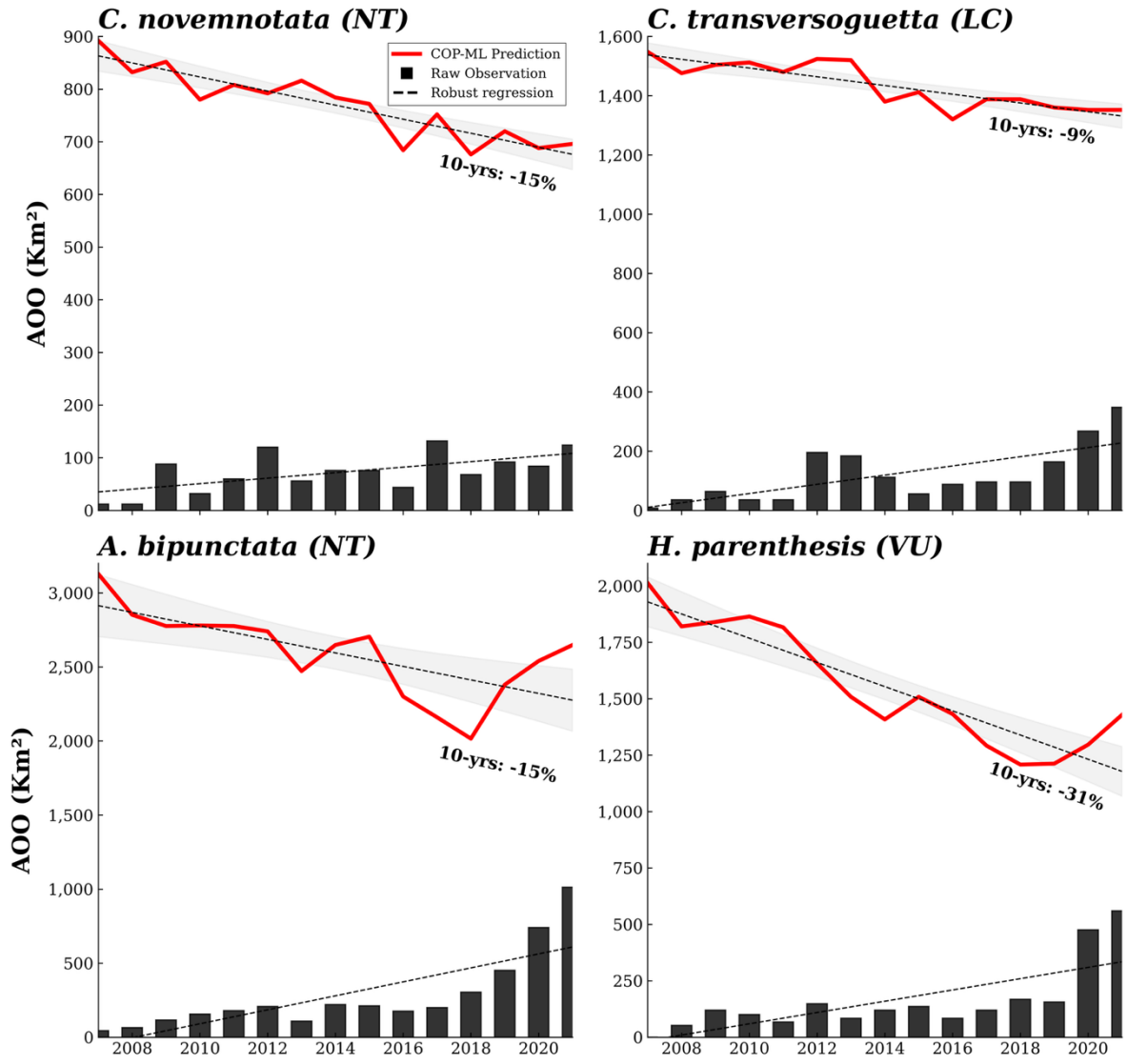


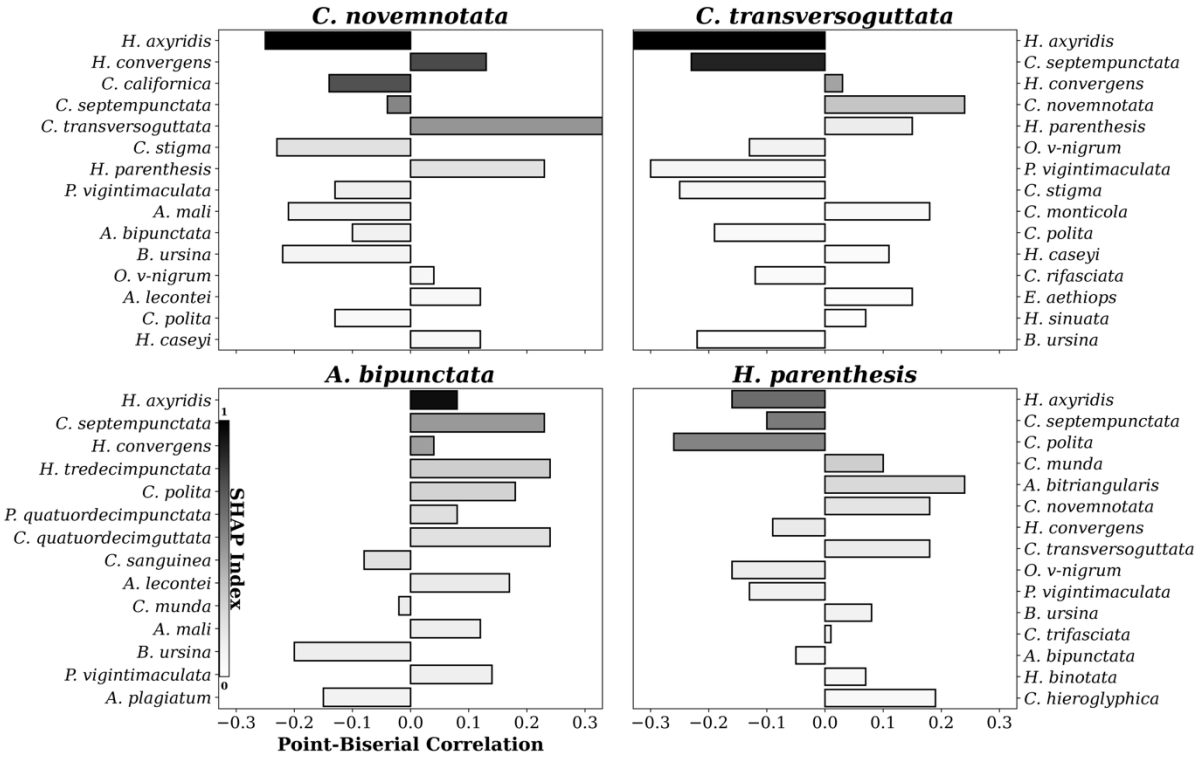
Figure 3. Annual occupancy of *Hippodamia parenthesis* (2007–2021) in the USA and Canada; Left, co-occurrence pattern model predictions at sites with prior records of the species; Right, raw observations from the compiled dataset; Dots, occupancy at previously recorded coordinates in the compiled dataset; color gradients, density of occupancy within each state or province.



834

835 **Figure 4.** Area of Occupancy (AOO) of four target ladybugs across the USA and Canada
836 (2007–2021); red lines, annual model predictions; dashed lines, robust regression fits showing
837 declines with 95% confidence intervals; IUCN Red List categories, derived from 10-year
838 reduction rates; bars, raw observations showing increases.

839



840

841 **Figure 5.** Predictor importance and correlation of co-occurrence pattern variables with target
842 species' occupancy (presence = 0, absence = 1); y-axis, SHapley Additive exPlanations (SHAP)
843 indices quantifying variable contributions in each model; x-axis, point-biserial correlation
844 between target occupancy and each variable species' occurrence within 18 km radius.

845

Appendix

Table S1. Ladybug records used were compiled from seven digital platforms (three citizen science, one museum collection, and three metadata sources); Regional abbreviations: AK = Alaska, HI = Hawaii, MB = Manitoba, ON = Ontario, SK = Saskatchewan, BC = British Columbia, AB = Alberta, QC = Quebec.

Source	Size	Type	Data Download & Refinement Criteria
Lost Ladybug Project	32,905	Citizen science	• Years 2007-2021 • U.S. (excluding AK & HI) and Canadian provinces (MB, ON, SK, BC, AB, QC) • Positional accuracy < 1 km (if applicable) • Species level • Adult records or images • Drop duplicates at year-GPS-species
iNaturalist	197,990	Citizen science	
bugGuide.Net	27,018	Citizen science	
GBIF	143,000	Metadata source	
BISON	109,834	Metadata source	
IdigBio	99,723	Metadata source	
NCSU Insect Museum	5,425	Institute	**Final Dataset:** 188,644

Table S2. ANOSIM statistics assess differences in co-occurrence pattern variables across data groups in the generalization test sets.

Species	Citizen science vs Institutions		Post-year (train) vs Pre-year (test)				Pre-year (train) vs Post-year (test)			
	R-value	*p*-value	Test period	Train (% of presence)	*R*-value	*p*-value	Test period	Train (% of presence)	*R*-value	*p*-value
C. transversoguttata	0.22	0.001	after 2019	65%	0.03	0.001	before 2014	70%	0.07	0.001
C. novemnotata	0.06	0.001	after 2018	70%	0.06	0.001	before 2013	72%	0.02	0.001
H. parenthesis	0.12	0.001	after 2020	74%	0.07	0.001	before 2014	74%	0.05	0.001
A. bipunctata	0.05	0.001	after 2019+	55%	0.02	0.001	before 2017	67%	0.02	0.001
Absence datapoints	0.06	0.001			0.03	0.001			0.01	0.001

+To assess COP model generalization, the training period for *A. bipunctata*—the species with most presence records—was reduced to extend the testing period.

Table S3. Results of regression estimates, diagnostic tests, and 2012-2021 reduction rates (OLS: ordinary least squares regression, Huber: robust regression).

	C. novemnotata	*C. transversoguetta*	*A. bipunctata*	*H. parenthesis*
B (OLS)	-13.36	-14.73	-45.59	-53.63
R^2 (OLS)	0.818	0.733	0.500	0.834
p (OLS)	0.0000****	0.0000****	0.0032***	0.0000****
95% CI (OLS)	-17.13, -9.59	-20.06, -9.4	-72.91, -18.26	-67.97, -39.29
Reduction (10-yr, OLS)	-15%	-9%	-15%	-29%
Breusch-Pagan *p*	0.9001	0.7992	0.0226*	0.0804
White *p*	0.7238	0.1968	0.0402*	0.0736
p (Robust SE)	0.0000****	0.0000****	0.0078*	0.0000****
95% CI (Robust SE)	-16.77, -9.95	-18.14, -11.31	-79.18, -11.99	-72.55, -34.71
Max Cook's Distance	0.2056	0.2028	0.6505	1.0651
B (Huber)	-13.00	-14.38	-45.98	-56.24
R^2 (Huber)	0.811	0.731	0.500	0.831
95% CI (Huber)	-16.42, -9.58	-17.39, -11.38	-72.32, -19.65	-68.23, -44.26
Reduction (10-yr, Huber)	-15%	-9%	-15%	-31%

($p^* < 0.05$, $p^{**} < 0.05$, $p^{***} < 0.005$, $p^{****} < 0.0005$)

Table S4. Predicted distribution trends (2007–2021) and corresponding IUCN Red List categories for four target species, based on reductions in area of occupancy (AOO) and extent of occurrence (EOO).

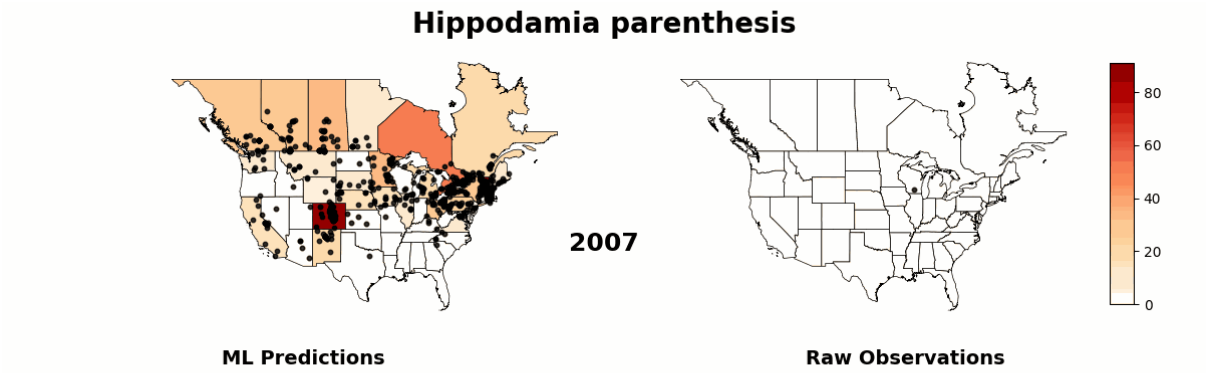
Species	Reduction in 10-yrs	IUCN category	AOO (km²)		EOO (km²)	
			2007	2021	2007	2021
H. parenthesis	31%	VU	1,548	1,352	8,450,469	7,749,070
A. bipunctata	15%	NT	3,128	2,648	11,538,691	10,817,443
C. novemnotata	15%	NT	2,012	1,428	5,480,067	5,399,901
C. transversoguttata	9%	LC	892	696	9,820,525	9,146,848

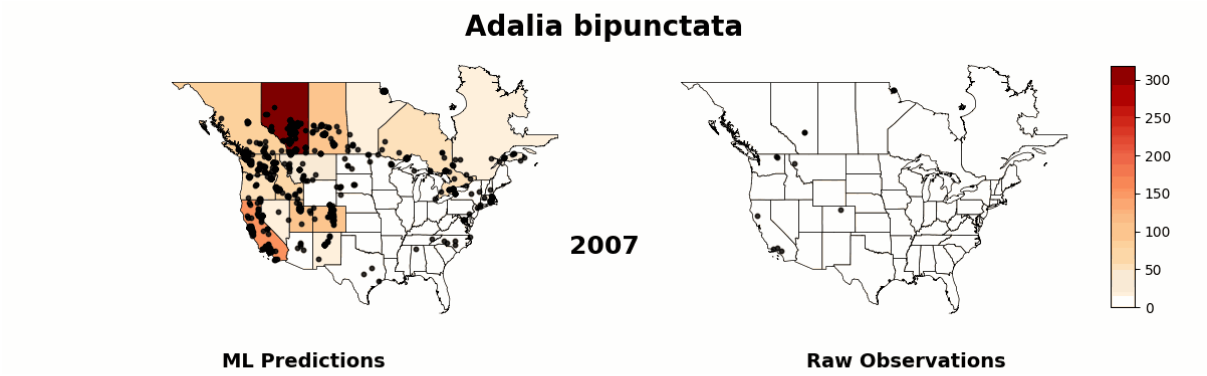
865 **Table S5.** Multiple linear regression (OLS) results assessing the effects of year and annual
866 citizen science data volume on ML-predicted annual area of occupancy (AOO) for four target
867 species. (**p < 0.05, **p < 0.005, ***p < 0.0005*).

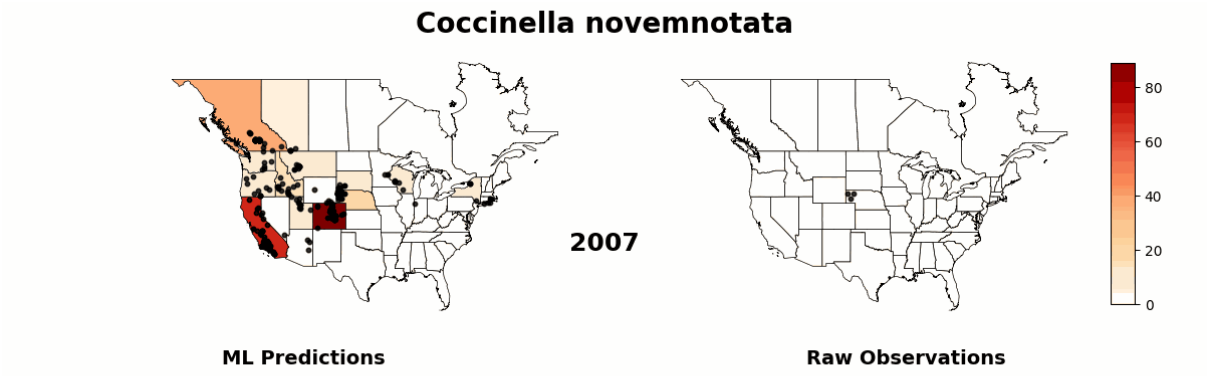
		C. novemnotata	*C. transversoguttata*	*H. parenthesis*	*A. bipunctata*
***F*-statistic** (DF Model, DF Residual)		12.96** (4, 10)	7.80** (4, 10)	34.72*** (4, 10)	12.86** (4, 10)
R²		0.83	0.76	0.93	0.84
		***B* coefficient (± *SE*)**			
Intercept		32224.9** ± *7617.4*	36645.9* ± *10517.8*	149163.4*** ±*18885.1*	160044.6*** ± *32274.6*
95% CI	Upper	49197.4	60081.1	191242.0	231956.8
	Lower	15252.3	13210.7	107084.8	88132.4
Year		-15.6** ± *3.8*	-17.5* ± *5.2*	-73.2*** ± *9.4*	-78.2*** ± *16.1*
95% CI	Upper	-7.1	-5.8	-52.3	-42.2
	Lower	-24.1	-29.2	-94.3	-114.0
Lost Ladybug Project		0.0003 ± *0.0236*	0.0221 ± *0.0326*	0.0870 ± *0.0586*	0.2587* ± *0.1001*
iNaturalist		0.0008 ± *0.0013*	0.0012 ± *0.0018*	0.0061 ± *0.0032*	0.0105 ± *0.0055*
bugGuide.Net		0.0680 ± *0.2646*	-0.0031 ± *0.3653*	-0.7286 ± *0.6560*	-1.9637 ± *1.1212*

868

869







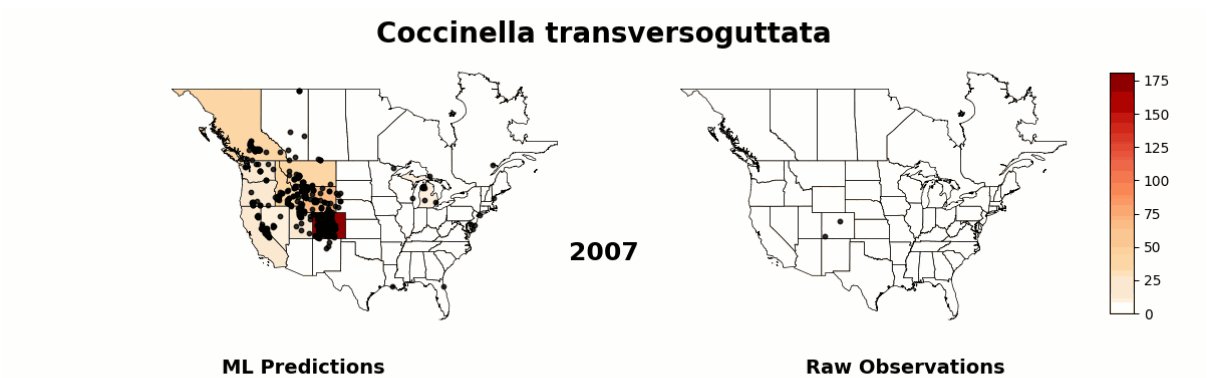


Figure S6. Annual occupancy maps (2007–2021) for each species; Left maps, occupancy predicted by co-occurrence–based models. Right maps, occupancy based on reported observations; Heatmap colors, the number of occupied coordinates per state, with temporal changes in color intensity reflecting shifts in occupancy (Active figures are available at: https://figshare.com/s/17cef8ef530f0a4f7b99).