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

HY Chung[1], DK Lee[2] and JE Losey[1*]

[1]*Department of Entomology at Cornell University, Ithaca, NY 14853;* [2]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, Structural Bias, Temporal Bias, Adventive Species, Native Ladybug Decline

WORD COUNTS

4,922

ACKNOWLEDGEMENTS

We extend our gratitude to Laura Melissa Guzman, Assistant Professor in the Department of Entomology at Cornell University, and Bongwon Suh, Associate Professor, and Jaeyoun You, Ph.D. Candidate in the Department of Intelligence and Information at Seoul National University, for their thoughtful reviews and comments.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA ACCESSIBILITY

Available form https://figshare.com/s/36131cf2516dc300e80a?file=54689660

ORCID

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

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

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

Abstract

Fragmented surveys and limited monitoring exclude most invertebrate species from conservation policy. We present a framework that generates annual occupancy predictions using species distribution models (SDMs) to reconstruct missing trends—not to extrapolate trends, but to fill them in (the fill-in approach). Instead of filtering poor-data regions and years or relying on static environmental variables, we use co-occurrence patterns (COP) to capture year-to-year shifts in species assemblages, to enable temporal prediction across all recorded habitats using sparse, presence-only datasets from multiple sources. 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 test splits, structurally distinct sources, and temporally divided periods. This indicates that annual predictions were robust to temporal bias arising from varying data volume and source composition, as supported by negligible effects in multiple regression. Predicted 10-year declines (9–31%) closely aligned with independent long-term regional monitoring data, operationalizing IUCN Red List classifications (from "Least Concern" to "Vulnerable") in the absence of standardized monitoring. By translating fragmented observations—primarily from citizen science—into standardized annual trend estimates, the fill-in approach extends extinction risk assessment to data-deficient taxa long excluded from conservation frameworks.

1 Introduction

Most invertebrate species—despite sharp declines (Montgomery et al., 2020)—remain invisible to conservation action, not because they are safe, but because they are silent in the data. Traditional approaches require revisits, effort-standardization, or abundance metrics—criteria that vast portions of invertebrate data simply fail to meet (Harvey et al., 2020). Without new tools to extract temporal signals from sparse, unstructured observations, these species will continue to decline unmeasured and unprotected.

Quantifying distribution trends for invertebrates remains a persistent methodological challenge, contributing to their significant underrepresentation in global conservation frameworks (Montgomery et al., 2020; Jönsson et al., 2021). Standardized monitoring programs are scarce or limited in scope (Estes et al., 2018; Bayraktarov et al., 2019), while available records frequently violate core statistical assumptions required by current trend detection models—ranging from ordinal abundance (e.g., Newson et al., 2015; Inamine et al., 2016; Schultz et al., 2017; Martín et al., 2021) and checklist-based data (Walker & Taylor, 2017; LeCroy et al., 2020) to survey effort metrics (e.g., Szabo et al., 2010; Isaac et al., 2014; Kamp et al., 2016; Horns et al., 2018; Fink et al., 2020) and repeat visits (e.g., MacKenzie et al., 2002; 2006; Kéry et al., 2010; van Strien et al., 2013; Altwegg & Nichols, 2019).

Opportunistic observations dominate in under-monitored taxa (Kissling et al., 2018), but introduce pronounced temporal and spatial biases (Isaac et al., 2014; Guzman et al., 2021, Larsen and Shirey 2021). While some strategies attempt to mitigate this through 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), such methods further reduce already sparse datasets, often eliminating precisely the regions and species of greatest conservation concern. As a result, the majority of invertebrate taxa remain excluded not for lack of ecological relevance, but for lack of usable data structures.

One untested strategy for bridging monitoring gaps is to reconstruct annual occupancy trajectories at historical locations using species distribution models (SDMs). Such models predict occupancy in unsampled areas from known presences (Olden et al., 2008; Zimmermann et al., 2010; Franklin, 2013), and have shown robust performance even under low data volumes (Hernandez et al., 2006; Wisz et al., 2008) or presence-only conditions (Robinson et al., 2018; Radomski et al., 2022).

However, most ML-based SDM applications focus on long-term range shifts, relying on static environmental predictors (Tingley & Beissinger 2009; Svancara et al., 2019), with rare attempts at annual trend detection, most of which depend on structured data and intensive thinning (Fink et al., 2020; Svancara et al., 2019), making them impractical for under-recorded invertebrates. Similarly, occupancy modeling approaches designed for trend detection demand strict revisit protocols and high spatiotemporal density (MacKenzie et al., 2002; Altwegg & Nichols, 2019)—e.g., in North America, ≥10,000 km² grids over ≥10-year intervals to achieve sufficient data density for bees and dragonflies (Soroye et al., 2020; Jackson et al., 2022). These data thresholds far exceed what is available for most under-recorded species, especially those under consideration in IUCN Red List assessments, which require trend estimates over a recent decadal timeframe (IUCN, 2024).

In short, traditional approaches—whether predictive or inferential—remain structurally inaccessible to the taxa most in need of conservation insight.

This study uses Co-occurrence-pattern predictors (COP) to generate fine-scale temporal predictions. COP describe the composition of nearby species within a radius (see Section 2.4). Prior studies suggest that these variables can embed species interactions and environmental associations (Pollock et al., 2014; Kissling et al., 2012). We propose that COP may also detect habitat and biotic changes more quickly than static environmental predictors, making them suitable for fine temporal scale, year-to-year occupancy prediction.

This study evaluates whether ML-based SDMs with COP variables (COP-ML) can generate accurate annual occupancy predictions from sparse, presence-only, multi-source datasets typical of under-recorded species, while remaining robust to temporal and structural biases.

To test for temporal bias, we assess whether models trained on one period generalize to others (Martínez-Minaya et al., 2018). Such generalization is essential (Willig et al., 2019; Guzman et al., 2021), as opportunistic citizen science (CS) data—now a major source of records (Knape et al., 2022)—is temporally uneven, often concentrated in recent years (Geldmann et al., 2016).

To test for structural bias, we evaluate whether models trained on one type of survey data can predict others. Multi-source integration is often necessary for rare species (Fletcher et al., 2019; Miller et al., 2019; Isaac et al., 2020), but inconsistent methods, even within CS platforms (Gardiner et al., 2012), can introduce systematic errors (Cheney et al., 2013). If models cannot accommodate such variation, observed changes may reflect protocol shifts rather than biological trends (Pagel et al., 2014; Knape et al., 2022).

This study evaluates three hypotheses: (1) Can ML-based classifiers distinguish target species' presence or absence using annual COP variables? (2) Can annual COP training enable generalization across survey methods? (3) Can this approach enable generalization across time periods? Then using COP ML's predictions, we reconstructed annual occupancy trajectories (2007–2021) for four native North American ladybugs and quantified 10-year declines to assign IUCN categories without structured monitoring.

2 Material and Methods

We assessed if COP variables could accurately predict annual species occupancy while mitigating temporal and structural biases without extensive filtering. Using six performance metrics, we evaluated COP-ML predictions of presence-absence across three scenarios: (1) predicting one survey group's data (e.g., institutions vs. citizen science) from another, (2) forecasting another time period using data from earlier or later years, (3) estimating 30% of the entire dataset from the remaining 70%. COP-ML, trained on the full dataset, then predicted annual occupancy from 2007-2021, enabling consistent comparisons of annual distributions for reduction rates and extinction risks.

2.1 Target species

Four native ladybug species—*Coccinella novemnotata*, *Coccinella transversoguttata*, *Adalia bipunctata*, and *Hippodamia parenthesis*—once dominated North American ladybug communities, thriving across diverse habitats and prey types (Losey 2007; 2012). Since the mid-1980s, their relative abundance in collections has dropped to 1/110–1/20 of former levels (Harmon et al., 2007) due to competition with adventive species *Coccinella septempunctata* and *Harmonia axyridis* (Wheeler & Hoebeke 1995; Harmon et al., 2007). These introduced species now dominate and reshape traditional communities, reducing diversity and abundance continent-wide (Petersen & Losey, 2024). Estimating reduction rates and extinction risks is challenging due to the natives' currently low density and wide distribution (Wheeler & Hoebeke, 1995; Hesler et al., 2004; Harmon et al., 2007), requiring integration of multi-source data across periods, regions, and methods while addressing inherent biases.

2.2 Occurrence data

Ladybug records were compiled from multiple sources: three CS platforms, a museum collection website, and three metadata platforms (Appendix Table S1). Two CS platforms, iNaturalist and bugGuide.Net, enabled user-identifications, while The Lost Ladybug Project relied on experts. We further verified target species identifications from iNaturalist and bugGuide.Net.

To assess how COP addresses biases, we applied minimal preprocessing. Data were restricted to the U.S. (excluding Alaska and Hawaii) and parts of Canada (Manitoba, Ontario, Saskatchewan, British Columbia, Alberta, Quebec) from 2007 to 2021, using only adult forms identified to species level. GPS accuracy, if available (89% of data), was limited to 1 km. Duplicates matching species-year-GPS were removed. Then, descriptive statistics are applied to reveal temporal and structural inconsistencies.

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* labeled as 'presence.'

2.3 Pseudo-absence

When explicit absence records are unavailable, presence records of other species serve as pseudo-absence points, typically with GPS locations randomly sampled from all other species in a pool (Robinson et al., 2018). In this study, we used GPS points of adventive species *C. septempunctata* and *H. axyridis* for two reasons. First, these species exclusively compete with the target species, and their presence within an 18 km radius—without target species— was assumed to represent logical absence and a reshaped COP after local extinction. Second, their dominance, 61% of our dataset, means conventional random sampling would still largely select these species, ensuring minimal methodological deviation.

We pooled 10,000 pseudo-absence points, selecting them from states or provinces proportional to the four target species' regional presence. Omitting the state ratio rule improved accuracy, but variable analysis showed an over-reliance on geographically specific variables, like *Coleomegilla maculata*, concentrated eastward. For our goal of predicting temporal changes, we prioritized biological interactions, such as competition, over static geographic distributions and introduced the matched state ratio treatment. We then randomly subsampled this pool multiple times for training and testing, labeling them as 'absence.'

2.4 Variables

Direct and indirect competition shapes ladybug assemblages, with adventive species dominance driving niche differentiation (Petersen & Losey, 2024) and avoidance behaviors in native species (Elliott et al., 1996; Hesler & Kieckhefer 2008; Mukwevho et al., 2017). We represented COP using the annual number of species records within an 18 km radius of presence and absence points, a distance based on typical ladybug dispersal ability (the exact number came from Jeffries et al., 2013; COSEWIC, 2016a; 2016b). For instance, most ladybugs are predators with high mobility (ex. *H. axyridis*, 442 km/year; McCorquodale, 1998) and active foraging across habitats (Woltz & Landis, 2013). Furthermore, we tested multiple radii (10–27 km) and selected 18 km as the smallest distance with sufficient data density, beyond which performance gains were marginal and ecological interpretability declined. To avoid self-guidance, we excluded target species' counts from their own variables. Variable's counts were min-max scaled by each species and year combination to correct for temporal and over-report variations in observation efforts. To reduce distributional bias during Min-Max scaling, outliers beyond the 1.5×IQR range from the 25th and 75th percentiles were removed.

We excluded environmental variables to prevent multicollinearity with COP variables (Kissling et al., 2012). From co-occurrence species, we retained 85 with at least 30 co-occurrences, excluding unidentified 'sp.' Multiple forward regressions ($p < 0.05$) selected predictive variables, with variance inflation factors (< 10) ensuring minimal multicollinearity. We ranked the top 15 key variables using SHapley Additive exPlanations (SHAP) values, which assess variable importance in model predictions. To examine relationships between variables and target species, we calculated average Point-Biserial Correlations by resampling pseudo-absence points 50 times to match presence record counts.

2.5 Development and characterization of models

We implemented the XGBoost Classifier (xgboost package) in Python, an ensemble method using gradient boosting trees to predict binary presence-absence (Chen & Guestrin, 2016). Default parameters were adjusted only for objective='binary:logistic' and n_estimators=1000 to optimize performance and regularization.

We applied a 7:3 train-test split ratio where applicable (see Section 2.5.1). For each test scenario, we balanced presence and absence at a 5:5 ratio by undersampling pseudo-absence points across 50 independent runs. Training and testing were then randomly split within these balanced datasets, yielding 2,500 unique iterations (50 splits × 50 subsamples).

We assessed model performance in annual occupancy prediction using six metrics: Accuracy (correct response rate), Kappa (considering default chance of true response; Cohen, 1960), Recall (true positive rate), and Precision (positive predictive rate) to measure ability in binary presence-absence predictions, plus Brier score (mean squared discrepancy; Brier, 1950) and AUC (class ranking; Fielding & Bell, 1997) for probability quality.

2.5.1 Generalization

Generalization tests evaluate a model's ability to predict data distinctive 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 our approach could generalize across structurally or temporally distinct data pools.

(1) Structural Generalization: We trained models on opportunistic CS datasets (LLP, iNaturalist, bugGuide.Net) to predict institutional datasets from 28 institutes, testing generalizability across survey types. COP differences between them were assessed using ANOSIM with Manhattan distance (Appendix Table S2). Presence records comprised 280 opportunistic 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, models trained on other sources (mean efficiency = 1.3) predicted LLP data (mean efficiency = 6.6), which emphasizes rare species monitoring; efficiency reflects the ratio of target species to total observations.

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

2.5.2 Evaluation

To assess COP-ML's annual prediction performance, we trained models using 70% of presence data and an equal number of pseudo-absence points, testing on the remaining 30%.

2.6 Prediction on annual distributions and reduction rates

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

(1) Prediction: We developed models as in 2.5, training them on all available presence data to improve prediction accuracy (Fielding & Bell, 1997; Rencher, 1995). A GPS point was deemed occupied in a given year if more than half of 2,500 models (50 train-test splits × 50 pseudo-absence subsamples) concurred.

(2) Analysis: We evaluated distribution trends using IUCN Red List Criterion A, based on changes in Area of Occupancy (AOO) and Extent of Occurrence (EOO). AOO, calculated as 4

km^2 grid cells occupied by a species, reflects occupancy extent and population size (IUCN, 2024). EOO, the polygon enclosing all known occurrences, indicates risk dispersion across a species' range (IUCN, 2024). For Criterion A, we fitted a linear regression to predicted AOO from 2007 to 2021 and estimated the most 10-year decline (2012–2021) from it, assuming these reflect population trends (IUCN, 2024). Heteroskedasticity was assessed with Breusch–Pagan and White tests, and influential outliers were identified using Cook's distance. Robust standard errors (HC3) were applied to account for heteroskedasticity and assess trend significance, followed by robust regression to derive the final AOO decline trend. Given that robust regression can exclude extreme values with abrupt changes which may reflect true trends, we also conducted ordinary least squares (OLS) regression to compare statistical estimates and improve the reliability of trend interpretation.

(3) Validation: To confirm that ML-predicted AOO changes reflect consistent temporal trends despite varying data availability, we used multiple linear regression with time (year) and annual CS source volumes as predictors.

3 Results

3.1 Biases in multi-source data

Our multi-source dataset exhibited structural and temporal biases. Structural bias, stemming from varying efforts and methods across sources (Figure 1), was evident in differing efficiencies for detecting target species. Institutional data (3.5% of the total dataset) recorded target species at 2.79 times the density of opportunistic data (96.5%). Even among citizen science platforms, LLP (5%, mean efficiency = 6.6) outperformed iNaturalist (89%, mean efficiency = 1.1) by sixfold in density. Temporal bias arose from an exponential rise in annual observations (Figure 1), with data volume post-2014 exceeding pre-2014 levels by 9.61 times.

3.2 Structural and temporal generalization

We tested COP-ML's annual prediction effectiveness and its generalizability against dataset biases through structural and temporal generalization tests. All models achieved reliable performance, exceeding benchmarks: Accuracy > 0.70, AUC > 0.70 (Hosmer et al., 2013), Kappa > 0.40 (Landis & Koch, 1977), and Brier < 0.25 (Brier, 1950; Figure 2). Models trained on unstructured CS datasets accurately predicted presence-absence in institutional datasets. Similarly, training on lower-efficiency datasets to predict higher-efficiency datasets, plus forward and backward temporal generalizations, met or surpassed these standards.

In structural generalization, *C. transversoguttata* model performed the best (0.87, 0.94, 0.75, 0.11), outperforming others: *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) in Accuracy, AUC, Kappa, and Brier scores. ANOSIM revealed small COP dissimilarities (< 0.25, p = 0.001; Appendix Table S2) between CS and institutional records, with R-values of 0.22 (*C. transversoguttata*), 0.12 (*H. parenthesis*), 0.06 (*C. novemnotata*), 0.05 (*A. bipunctata*), and 0.06 (absence points).

For temporal generalization, *C. transversoguttata*, *C. novemnotata*, and *H. parenthesis* maintained consistent performance regardless of direction. Conversely, *A. bipunctata*'s backward performance dropped 7%, with Recall (true positives among actual positives) rising 2% and Precision (true positives among predicted positives) falling 13%. This indicates that models trained on recent data classified broader habitat conditions as occupied than were historically, suggesting its current occupancy may generalize beyond past habitat needs, unlike the other species. A supplementary analysis, expanding training data to 87% and limiting recent occupancy from test data (pre-2012), intensified this trend: Recall rose 11% and Precision fell 18%, while Accuracy remained unchanged.

3.3 Evaluation of the developed models

We assessed COP-ML classifiers, trained on 70% of the full multi-source dataset and tested on 30%, against established standards and found them practical. *C. transversoguttata* (510 presence points) showed the highest performance, followed by *C. novemnotata* (324), *A. bipunctata* (1,438), and *H. parenthesis* (742). Even the lowest-performing species exceeded satisfactory benchmarks: Accuracy, Precision, and Recall > 0.75 (excellent), AUC > 0.87 (outstanding), Kappa > 0.57 (substantial), and Brier < 0.15.

3.4 Predicted reduction rates and conservation status

Given COP-ML's practical performance in prior tests, we predicted annual occupancy to fill monitoring gaps across all historical observation points from 2007 to 2021, ensuring consistent temporal comparisons (Figure 3; Appendix Figure S1).

All species showed statistically significant AOO decline trends (p < 0.05 for OLS and robust SE). Heteroskedasticity in *A. bipunctata* and influential outliers in *A. bipunctata* and *H. parenthesis* were detected, but differences in OLS, robust SE, and robust regression estimates (B, R^2, CI) were small, confirming reliable decline trends (Appendix Table S3).

Predictions suggested three target species are threatened by continuous declines in North America (Appendix Table S4). Area of occupancy (AOO), an indicator of occupied area and indirectly population size (IUCN, 2024), declined across all four species from 2007 to 2021: *H. parenthesis* by 1,962 km^2, *A. bipunctata* by 584 km^2, and *C. novemnotata* and *C. transversoguttata* by 480 km^2 each. Per IUCN Red List Criterion A, 10-year reduction rates estimated *H. parenthesis* at 31% ("Vulnerable"), *A. bipunctata* at 15% ("Near Threatened"), *C. novemnotata* at 15% ("Near Threatened"), and *C. transversoguttata* at 9% ("Least Concern"; Figure 4).

Extent of occurrence (EOO), reflecting spatial risk dispersion (IUCN, 2024), declined most in *C. transversoguttata*. Despite its "Least Concern" status in this study, this species indicates reduced extinction resistance with ongoing population decline.

Multiple linear regression confirmed that time (year) significantly drove AOO declines across all species, while annual citizen science data volumes showed no evidence of statistically or practically meaningful effects (Appendix Table S5).

3.5 Variable importance and correlation

SHAP values and Point-Biserial correlations revealed positive interdependence among *C. novemnotata*, *C. transversoguttata*, and *H. parenthesis*, with their predicted presences linked in ML models (Figure 5). Conversely, *H. axyridis* and *C. septempunctata*, ranking as the most influential variables, showed negative correlations with these three species. *A. bipunctata*, however, exhibited a positive correlation with them, marking an exception. *H. convergence*, a native species with the third highest abundance in the dataset, also correlated positively with three natives, contributing significantly to all models except *H. parenthesis*.

4 Discussion

4.1 Rationale for estimated reduction rates

This study provides the first continent-wide estimates of decadal occupancy declines for *C. novemnotata*, *C. transversoguttata*, *A. bipunctata*, and *H. parenthesis* based on annual presence predictions. Earlier studies from the 1980s–1990s reported steep relative abundance declines—up to 95–99% (rescaled from Harmon et al., 2007)—whereas our more moderate reduction rates from 2007–2021 likely reflect the post-establishment phase of dominant adventive species.

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).

Subsequent regional studies suggest that native species 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 re-equilibration, 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, documenting 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 based on linear regressions of sticky-trap captures normalized by survey effort. Although limited to a single available site, these local declines align closely with our continent-wide estimates (31% and 15%, respectively), suggesting that the COP-based annual predictions provide ecologically realistic baselines for broader conservation assessments.

4.2 Interpretation of COP variables

Annual prediction accuracy likely reflects the extent to which COP variables encode dynamic ecological processes—such as species interactions and habitat turnover—beyond what static predictors like climate or land cover can capture. Despite attenuation bias from observational noise across large, heterogeneous sources—which likely suppressed effect sizes—the direction and relative influence of key COP predictors remained largely consistent with known ecological associations (Figure 5).

For instance, *C. novemnotata* and *C. transversoguttata* showed strong positive associations, consistent with overlapping habitat use and resource preferences (Hesler et al., 2009). In contrast, *H. axyridis* and *C. septempunctata*—the two most influential variables—were negatively associated with three target natives, as known patterns of competitive displacement (Wheeler & Hoebke, 1995; Harmon et al., 2007; Petersen & Losey, 2024).

By contrast, *A. bipunctata* showed positive associations with both adventive species, likely due to macro-scale aboreal habitat preference overlap with *H. axyridis* (Coderre et al., 1995; Koch, 2003; Omkar & Pervez, 2005; Hentley et al., 2016) and overlap with *C. septempunctata* in Europe where both are native (Honěk 1985; Nedvěd 1999). This pattern does not preclude competitive exclusion at finer spatial scales (e.g., <18 km), which may not be captured within the COP resolution used (Kajita et al., 2000; Kajita et al., 2006; Soares & Serpa, 2007).

Lower-ranked COP variables showed weaker links to known ecological interactions and may function primarily as spatial or environmental proxies. Their limited influence suggests that predictions were mainly driven by biologically meaningful co-occurrence patterns, rather than incidental spatial overlap.

4.3 Strength and limitation of COP

The robustness of COP models across periods and sources suggests that co-occurrence structures may encode latent ecological constraints—such as competitive exclusion or shared habitat filtering—that static environmental variables often fail to capture.

COP-based predictors would be effective when biotic interactions strongly shaped distributions. In our case, native ladybug communities were shaped by prolonged competition with adventive species (Harmon et al., 2007; Petersen & Losey, 2024), and COP variables, even from opportunistic data, reflected these patterns.

However, this strength is contingent on the temporal stability of species interactions (Tikhonov et al., 2017). COP-ML declined in performance when backcasting the distribution of *A. bipunctata*, a species whose habitat preferences were reported to have shifted due to post-invasion habitat compression (Bahlai et al., 2015). The model tended to overpredict past suitability—evidenced by higher recall than precision. This pattern suggests a limitation of COP-based models when underlying biotic interactions shift over time, as static co-occurrence relationships may no longer align with changing ecological realities.

Nevertheless, COP variables exhibited strong generalizability across heterogeneous datasets—ranging from open-ended citizen science to a targeted rare-species initiative—while maintaining minimal structural divergence (Appendix Table S2). This highlights their resilience to source variation—a critical property in the context of conservation modeling. Historically, skepticism toward unstructured data has limited its utility (Isaac & Pocock, 2015; Steen et al., 2019), even while multi-source integration becomes increasingly recognized as essential for data-deficient taxa (Miller et al., 2019; Isaac et al., 2020). Our results show that COP—driven by commonly recorded species—can indirectly reveal the distributions of rarer taxa, enhancing the conservation value of citizen science and leveraging its rapid growth in data volume.

This likely stems from the robustness of relational signals: co-occurrence patterns tend to be more resilient to sampling noise than marginal occurrence rates of individual species, which are often more sensitive to variation in effort or detection error (Tikhonov et al., 2017; Johnston et al., 2017). COP variables would leverage these dependencies to enable ecological inference even from opportunistic or sparse data. Our results were derived from a dataset incorporating over 78 institutions and projects—97% originating from citizen science and with minimal filtering applied. Given their consistency under such heterogeneous and unstructured conditions (Appendix Table S2), observed shifts in COP structure are more likely to reflect ecological change than artifacts of sampling noise. In this context, the robustness of COP variables implies more than resistance to bias; it provides empirical grounds to interpret persistent co-occurrence signals as evidence of underlying biotic structure.

4.4 fill-in approach with annual predictions

This study proposes a ‘fill-in’ approach, uniquely generating annual occupancy predictions to bridge monitoring gaps. By tracking occupancy changes across North American habitats since 2007, we evaluated extinction risk for four ladybug species under the IUCN Red List's recent 10-year population reduction criterion (Figure 3).

Traditional time-series methods often rely on a "filtering" strategy—retaining only well-monitored regions and thus narrowing the spatiotemporal scope of inference. Model-based approaches, by contrast, often require data-intensive population modeling (Fink et al., 2020) or repeated surveys—resources typically unavailable for under-recorded taxa. Our approach generates fine-scale temporal predictions from sparse, presence-only, multi-sourced datasets, benefitting these taxa.

Integrating annual predictive modeling with existing frameworks may enhance their applicability to data-deficient species. Occupancy models (OM), for instance, track temporal distribution shifts but require at least two revisits per period (Royle, 2006; Kamp et al., 2016; Outhwaite et al., 2018; Perkins-Taylor & Frey, 2020; Jha et al., 2022). Across North America, sufficient data density for bees and dragonflies required spatial and temporal resolutions as coarse as 10,000 km² and 10–20 years (Soroye et al., 2020; Jackson et al., 2022; Shirey et al., 2023). Rather than compromising resolution or re-sampling unsurveyed areas with additional costs (Xue et al., 2016, Tulloch et al., 2013), annual ML predictions can offer an efficient alternative.

However, incorporating ML-based predictions into occupancy modeling frameworks—as pseudo-observations—requires further validation to ensure statistical compatibility. While our models reliably tracked distributional changes, OM frameworks involve explicit modeling of detection and survey processes that must be reconciled with predicted data. Though OM advancements explored application of non-ideal data—e.g., assuming observation processes as random walks, using pseudo-absence instead of checklist absence, assuming opportunistic observations as revisit surveys (Outhwaite et al., 2018)—OM's fit with fill-in predictions is untested. Combining annual predictive modeling with existing methods holds significant potential in conservation, but it requires identifying appropriate integration strategies and evaluating their logical consistency, performance enhancements, and the validity and reliability of results.

The fill-in and filtering approaches can be complementary. In this study, we did not account for spatial bias or spatial autocorrelation, often addressed through spatial thinning—a widely used filtering method. Filtering, however, introduces trade-offs: it narrows the spatiotemporal scope of inference, excludes rare species, and undermines the reliability of absolute-scale assessments such as IUCN Red List categorizations. We deliberately avoided filtering for three reasons: first, our focus on rare species with limited data rendered filtering impractical. Second, we aimed to test whether our approach, designed for such species, could perform well without relying on filtering (see Section 3.2, 3.3). Third, our goal was to generate predictions across the full known range of each species. However, balancing the selection of high-quality data through filtering with the benefits of more training data may be critical. Future research should determine the optimal data filtering level to improve the accuracy and reliability of predictions, retain data volume, and reduce bias.

This study presents a scalable method to bridge monitoring gaps for data-deficient species, using sparse, presence-only records—largely from citizen science—to generate annual occupancy estimates. The 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 extinction risk signals can emerge even from unstructured datasets. By transforming fragmented observations into interpretable trends, the fill-in approach provides a practical pathway to extend extinction risk assessment and strengthen conservation decisions in the absence of standardized monitoring.

5. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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.

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

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

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

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

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

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

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

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

Hesler, L. S., Kieckhefer, R. W., & Catangui, M. A. (2004). Surveys and field observations of Harmonia axyridis and other Coccinellidae (Coleoptera) in eastern and central South Dakota. *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.*

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.

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., ... & 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Miller, D. A., 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.

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

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

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

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

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

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

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

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

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

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

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

Radomski, T., Beamer, D., Babineau, A., Wilson, C., Pechmann, J., & Kozak, K. H. (2022). Finding what you don't know: Testing SDM methods for poorly known species. *Diversity and Distributions*, **28(9)**, 1769-1780.

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

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

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

Schultz, C. B., Brown, L. M., Pelton, E., & Crone, E. E. (2017). Citizen science monitoring demonstrates dramatic declines of monarch butterflies in western North America. *Biological 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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

Figures and Tables

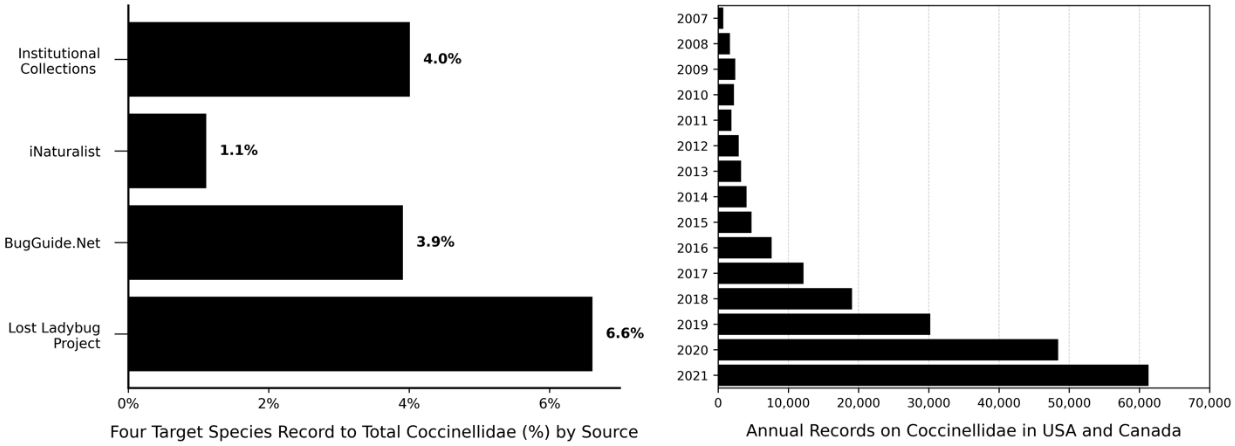

Figure 1. Left: Detection efficiency of four target coccinellid species in this study across data sources, highlighting structural inconsistencies among opportunistic citizen science platforms (The Lost Ladybug Project, iNaturalist, BugGuide.net) and institutional records (see Section 3.1). **Right**: Exponential increase (y = 758.89$e^{0.29x}$ (R^2 = 0.88)) in the annual number of coccinellid observations across the USA and Canada (2007–2021), showing a 9.61-fold rise post-2014 relative to pre-2014, highlighting temporal bias (see Section 3.1).

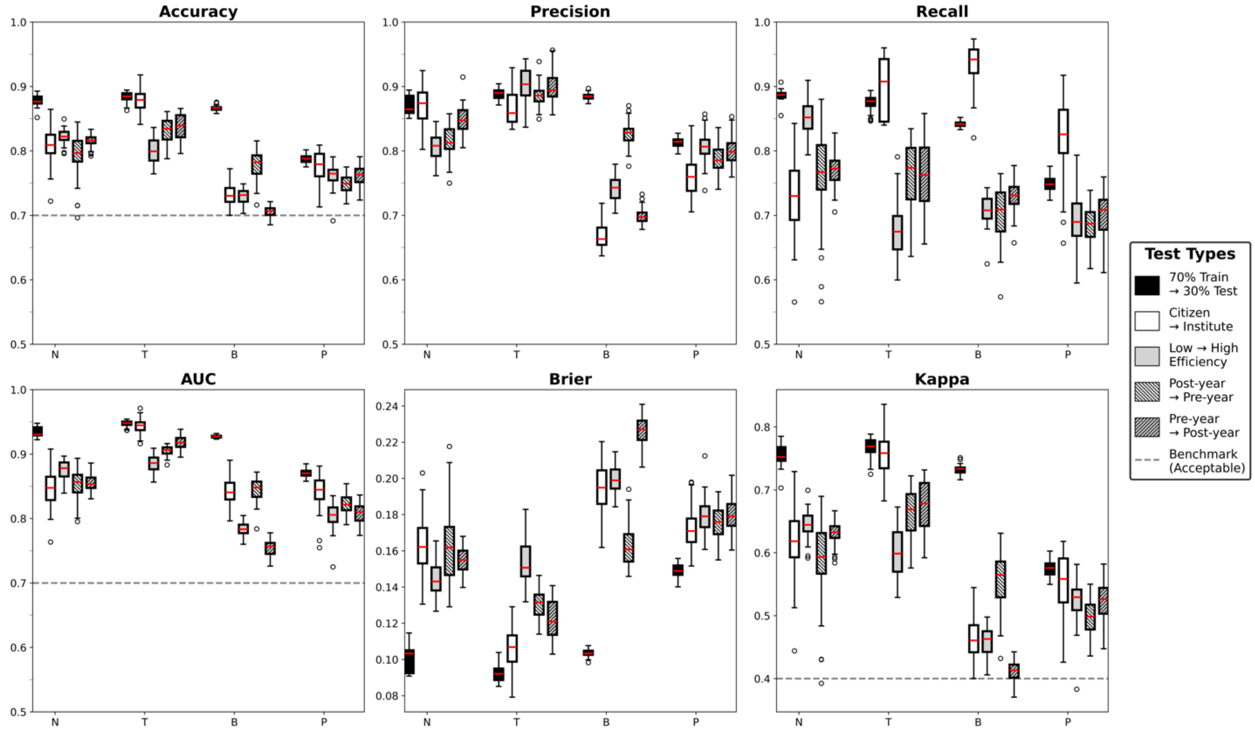

Figure 2. Performance of species distribution models using co-occurrence patterns for *C. novemnotata* (N), *C. transversoguttata* (T), *A. bipunctata* (B), and *H. parenthesis* (P) across five data settings (see Section 3.2, 3.3). Black plots represent a 70:30 train-test split. Structural generalization tests include training on citizen science and testing on institutional data (white plots), or training on low-efficiency sources (1.3% target species detection) and testing on a high-efficiency rare species monitoring source (6.6%; gray plots). Temporal generalization tests use post-2007 training with pre-2021 testing (left diagonal hatch) or the reverse (right diagonal hatch), with a cutoff year at approximately 70% training data coverage. The red line indicates the mean performance across 2,500 training iterations, with a 95% confidence interval. All models exceeded acceptable performance benchmarks: Accuracy > 0.70, AUC > 0.70, Kappa > 0.40, and Brier < 0.25 (Hosmer et al., 2013; Landis and Koch, 1977; Brier, 1950).

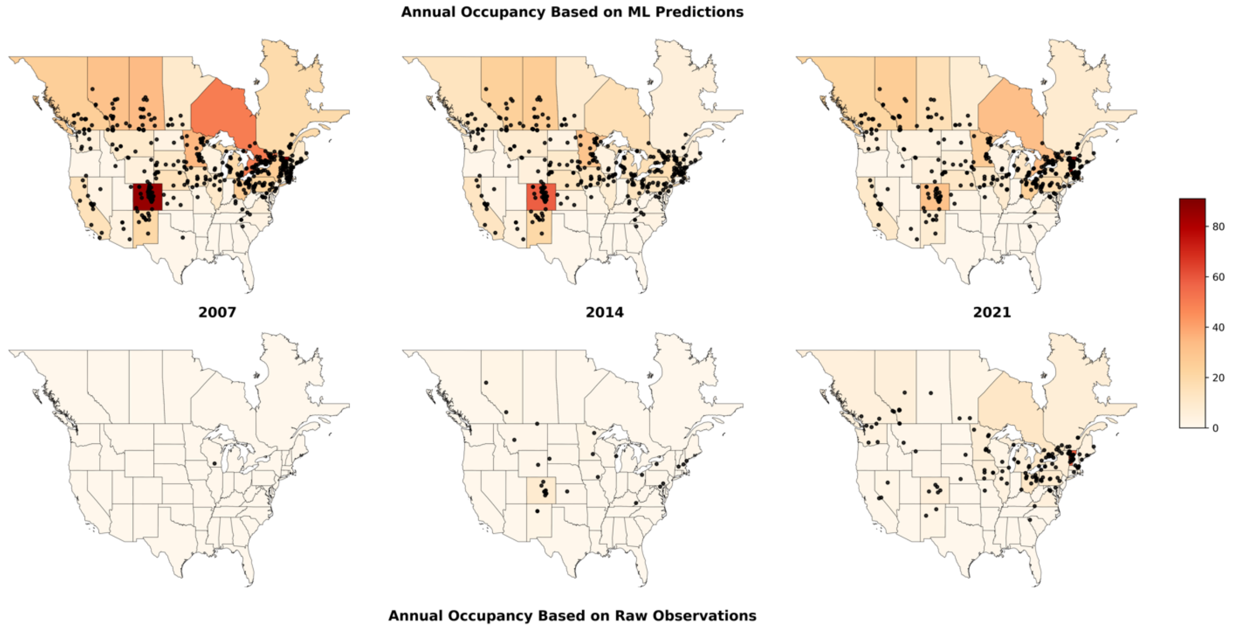
Annual Occupancy Based on ML Predictions
2007
2014
2021
80
60
40
20
0
Annual Occupancy Based on Raw Observations

Figure 3. Annual occupancy of *Hippodamia parenthesis* (2007–2021) in the USA and Canada, comparing co-occurrence-based model predictions (upper maps) with raw observations (lower maps). Dots show occupied locations within each state, with color gradients to represent state-level occupancy changes over time (see Section 3.4).

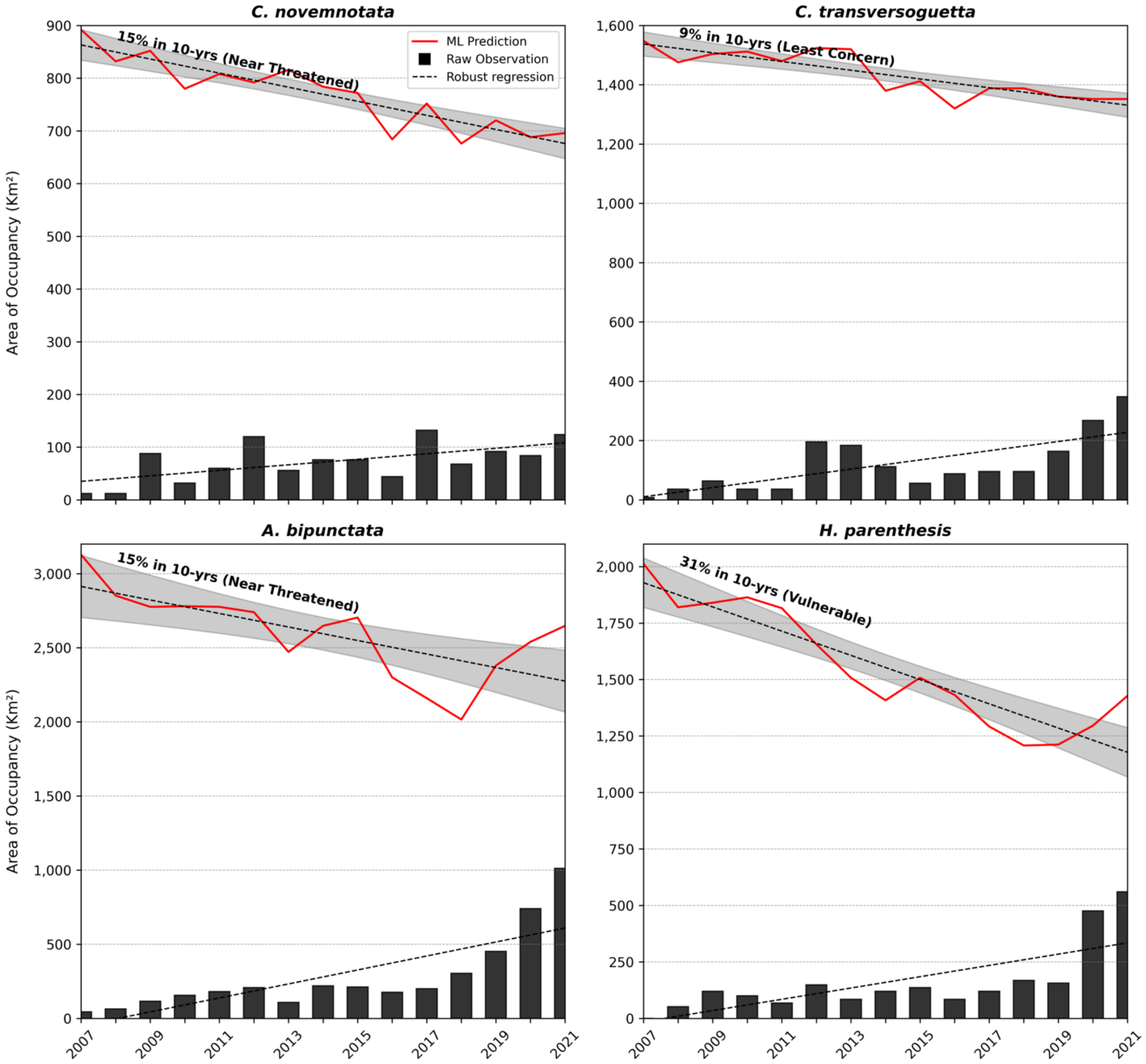

Figure 4. Area of occupancy for four target coccinellid species in the USA and Canada (2007–2021), with annual predictions (red lines) showing declines, while raw observations (bars) suggest increases due to temporal bias. Dashed lines show robust regression trends with 95% confidence intervals, with IUCN Red List categories based on 10-year reduction rates (see Section 3.4).

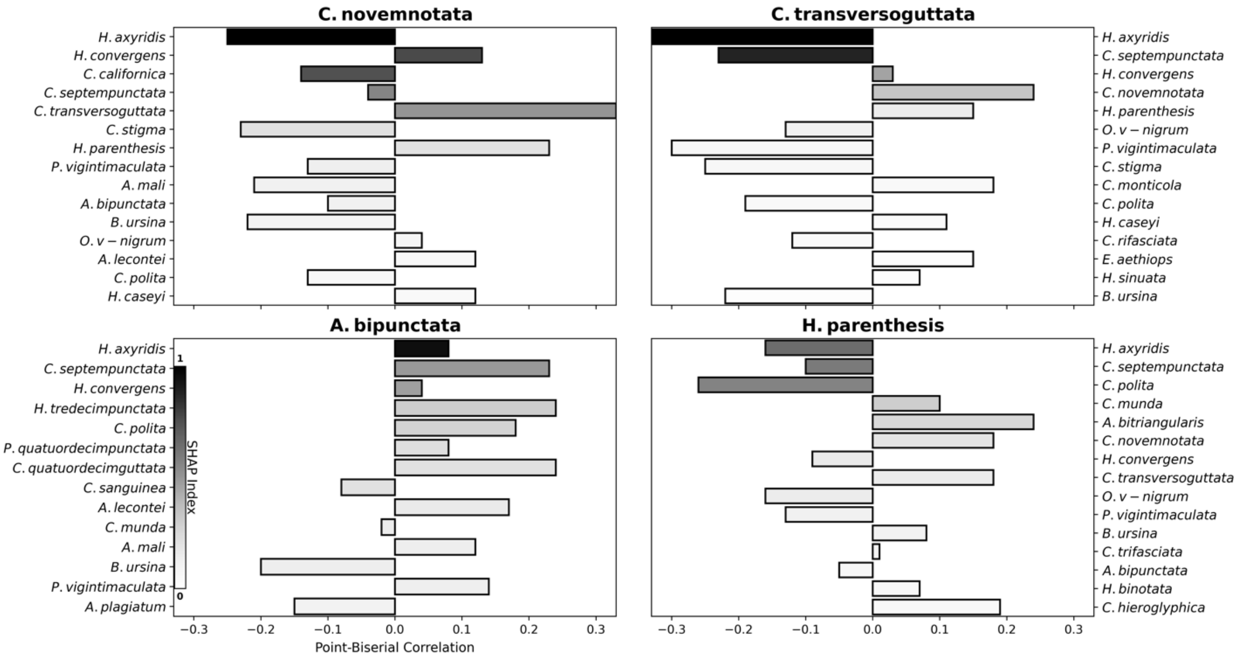

Figure 5. Variable importance is ranked by the SHapley Additive exPlanations index (y-axis), while the Point-Biserial Correlation (x-axis) quantifies the association between variables and the presence of target species (see Section 3.5).

Appendix 1

Table S1. Occurrence records of coccinellid species from seven digital platforms (three citizen science, one museum collection, and three metadata sources) used in this study (see Section 2.2). Regional abbreviations: AK = Alaska, HI = Hawaii, MB = Manitoba, ON = Ontario, SK = Saskatchewan, BC = British Columbia, AB = Alberta, QC = Quebec.

Source	Size	Type	Data Download & Refinment 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 • Only 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 results assessing differences in co-occurrence patterns (COP) across data groups in the generalization test sets (see Section 2.5.1).

Species	Citizen science and institutional data		Post-year train, Pre-year test				Pre-year train, Post-year test			
	ANOSIM value	**p-value**	**Test year period**	**Train** (% of presence)	**ANOSIM value**	**p-value**	**Test year period**	**Train** (% of presence)	**ANOSIM 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
***R²* (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
***R²* (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 IUCN Red List status of four rare coccinellid species based on reductions in area of occupancy (AOO) and extent of occurrence (EOO; see Section 3.4).

Species	**Reduction in 10-yrs**	**IUCN status**	**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

Table S5. Multiple linear regression (OLS) results evaluating the effects of time (year) and annual data volume from citizen science sources on ML-predicted annual area of occupancy (AOO) for four target species (*p < 0.05, **p < 0.005, ***p < 0.0005; see Section 3.4).

		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)
***R*²**		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*

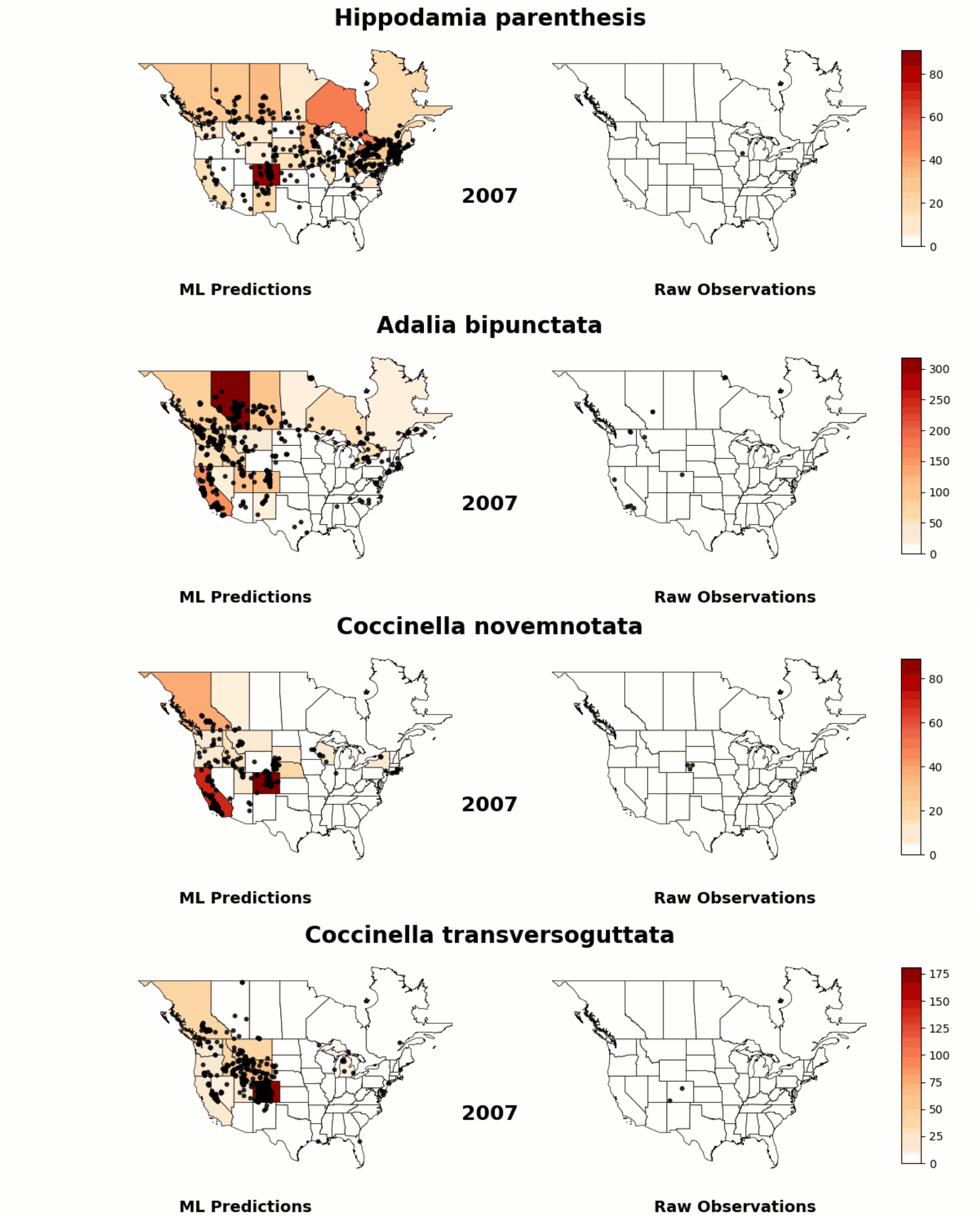

Figure S1. The maps depict the occupied coordinates of each species for each year from 2007 to 2021. The left maps show annual occupancy predicted by co-occurrence-based models, while the right maps are based solely on reported observations. The heatmap represents the number

of occupied coordinates per state, with color shifts over time indicating changes in occupancy. (Active figures are available at: https://figshare.com/s/17cef8ef530f0a4f7b99)