# Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science

GRACE J. DI CECCO, VIJAY BARVE, MICHAEL W. BELITZ, BRIAN J. STUCKY, ROBERT P. GURALNICK, AND ALLEN H. HURLBERT

*The availability of citizen science data has resulted in growing applications in biodiversity science. One widely used platform, iNaturalist, provides millions of digitally vouchered observations submitted by a global user base. These observation records include a date and a location but otherwise do not contain any information about the sampling process. As a result, sampling biases must be inferred from the data themselves. In the present article, we examine spatial and temporal biases in iNaturalist observations from the platform's launch in 2008 through the end of 2019. We also characterize user behavior on the platform in terms of individual activity level and taxonomic specialization. We found that, at the level of taxonomic class, the users typically specialized on a particular group, especially plants or insects, and rarely made observations of the same species twice. Biodiversity scientists should consider whether user behavior results in systematic biases in their analyses before using iNaturalist data.*

*Keywords: citizen science, iNaturalist, biodiversity*

**We are currently in a period of unprecedented** growth in the global availability of species occurrence records as a result of data collected through citizen science projects (sometimes known as *community science* or *participatory science*; Bonney et al. 2014, Brown and Williams 2019). Although many of these records come from structured or semistructured surveys that include information on the sampling effort (these records are particularly numerous for birds), the vast majority of observations for other taxonomic groups are from unstructured, opportunistic observations that simply link a taxonomic entity to a particular point in time and space (Welvaert and Caley 2016, Pocock et al. 2017, Kelling et al. 2019). The lack of formal survey procedures for contributing unstructured observations has meant a low barrier of entry for participation and has facilitated the accumulation of large amounts of data. These opportunistic biodiversity observations have been applied to a variety of ecological questions across taxonomic groups, including terrestrial vertebrates, invertebrates, and plants, as well as marine organisms (Follett and Strezov 2015).

One of the largest unstructured biodiversity survey projects spanning the globe is iNaturalist (www.inaturalist.org/), a joint initiative of the California Academy of Sciences and the National Geographic Society, which provides an online platform for recording and identifying observations of any species. Users upload an observation of an organism (typically a photo, although sound recordings are now permitted) and can propose an identification or receive suggestions from community members. iNaturalist also shares records that meet certain quality thresholds through the Global Biodiversity Information Facility (GBIF; www.gbif.org). The platform provides an online forum for individuals interested in documenting the organisms they encounter, providing taxonomic identifications of the observations of others, and participating in a community of fellow naturalists while producing data that can be used by scientists (www.inaturalist.org/pages/about).

The massive scale of data available on iNaturalist, over 56 million observations at the time of writing and roughly doubling each year, has led to a surge in research, making use of these data to address a variety of research questions. iNaturalist observations have been used to assess phenology, such as using photographs associated with observations to identify flowering duration (Li et al. 2020) and unusual

*BioScience* XX: 1–10. 
https://doi.org/10.1093/biosci/biab093

flowering events (Barve et al. 2020). For these applications, iNaturalist data enabled analyses at a broader spatial extent and provided clearer separation of cultivated and wild organisms than other commonly used phenology-monitoring data. In addition to phenological state, photographs of organisms may also contain useful information about organismal phenotype. Drury and colleagues (2019) used iNaturalist data to document geographic variation in wing phenotypes of two species of damselflies, testing previously suggested hypotheses regarding character displacement and the processes influencing trait evolution across landscapes.

iNaturalist observations also have increasingly useful applications in research focused on monitoring biodiversity, especially for species that are readily detectable and identifiable via photograph. For example, iNaturalist observations have been used to identify a threatened species of bumblebee that had not been reported in several decades in the Philippines (Wilson et al. 2020) and to track invasions of a mantis species in France (Moulin 2020) and ladybird beetles in Argentina (Werenkraut et al. 2020). Images including multiple organisms (e.g., flowers in pictures of bumblebees) may also be used to examine interactions between species (Gazdic and Groom 2019).

One of the most common research uses of iNaturalist data is the development of species distribution models, especially records that are included in GBIF (Heberling et al. 2021). iNaturalist records have been used to build species distribution models of plants (Chapman et al. 2019), reptiles and amphibians (Fourcade 2016), and other vertebrates and invertebrates (Heberling et al. 2021). iNaturalist records have also been used to characterize climatic tolerances of species in studies of shifting range limits (Chardon et al. 2015) and to supplement observations to evaluate the conservation status of a species of poison dart frog (Balaguera-Reina et al. 2019). Other biodiversity research applications include characterizing community composition over time (Rappaciuolo et al. 2021) and describing species tolerances of urban habitats (Callaghan et al. 2020). Because data from unstructured, opportunistic observations lack information about the sampling or reporting process, an understanding of user behavior and the data collection process must be inferred from the data themselves in order to mitigate any potential biases.

## Understanding biases in citizen science data

Fundamental to the success of platforms such as iNaturalist is an engaged and growing user base willing to volunteer their time to collect and identify biodiversity records. User observation patterns are critical to examine because they effectively determine how spatial, temporal, and taxonomic biases are structured. For example, because observations tend to be made in the volunteers' free time, a strong bias toward increased observations on weekends has been documented in bird citizen science observations (Courter et al. 2013). Some have argued for categorical grouping of users on the basis of behavior (Boakes et al. 2016), whereas others have suggested that the participants can be more appropriately viewed on a continuum of frequency and intensity of platform use (August et al. 2020). Efforts to measure user behavior and the associated sampling biases have often been focused on a few well-studied taxonomic groups. However, the power of augmenting traditional, structured sources of biodiversity data with unstructured observations is greatest in groups that are not well sampled, making it particularly useful and needed to investigate spatial and taxonomic biases and user behavior across a broad taxonomic scope. Such an empirical description of how users record observations on iNaturalist will better inform the usage of opportunistic data for biodiversity research.

We examined the full set of iNaturalist observations that had been uploaded since its founding in 2008 through the end of 2019 (over 31 million observations) to better understand how observers make use of iNaturalist. In this study, we describe spatial and temporal biases in where and when observations are collected. We also investigated the extent to which users specialize taxonomically in their observations and describe the most common types of taxonomic specialization. Our results have implications for biodiversity scientists seeking to use iNaturalist occurrence data in a way that accounts for biases in observation distribution and user behavior.

## The iNaturalist data set

iNaturalist data were downloaded using iNaturalist's web API via the R package "rinat" (Barve et al. 2020). The website started operations in March of 2008, and the data was downloaded for all years from 2008 to 2019 for a total of 31 million observations. Custom scripts were used to download the higher taxonomy data for each taxon included in the iNaturalist occurrence data. For all observations, we used the date of observation rather than the date of upload for analyses and the species identification associated with the observation at the time of download.

In each year from 2008 through 2019, we counted the total number of iNaturalist observations, as well as the number of unique iNaturalist users who made at least one observation (hereafter, "users") during that year. We also characterized intra-annual variation in the number of users and observations per week within a single year, 2018. To examine spatial patterns in iNaturalist observations, we calculated the number of observations per country and for observations within the conterminous United States, we used land-cover classification from the National Land Cover Database (NLCD) 2016 version (Yang et al. 2018) to calculate the density of sampling within land cover types.

We examined the completeness of iNaturalist's species-level coverage by comparing the number of species represented by iNaturalist observations for a given taxonomic class with the total number of described species within the taxonomic class as estimated by the Catalog of Life (COL; Roskov et al. 2020). We obtained species-per-class counts from the 1 September 2020 monthly edition of COL. We counted only

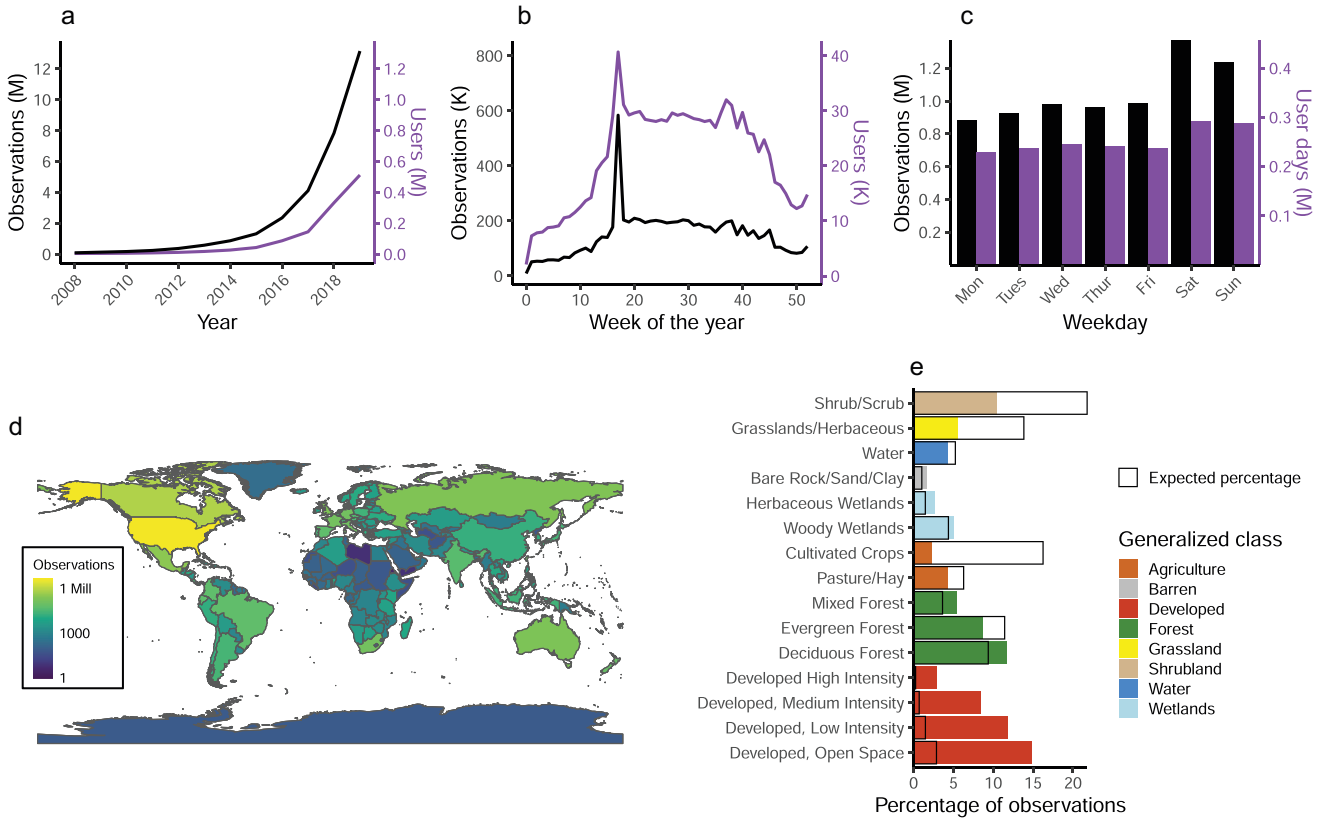


***Figure 1. Spatial and temporal patterns in iNaturalist observations submitted through 31 December 2019. (a) Growth in unique users making observations and observations over time. (b) Weekly number of observations submitted and unique users from 1 January 2018 to 31 December 2018. (c) Total number of observations submitted and user days per day of week from 1 January 2018 to 31 December 2018. (d) Global distribution of observations by country. (e) Land cover classification of observations made in the coterminous United States from the National Land Cover Database 2016 version; solid bars showing number of observations per land cover class, outlined empty bars showing expected percentage of pixels in that class across the coterminous United States.***

COL records with "taxonomicStatus" equal to "accepted name" or "provisionally accepted name" (e.g., synonyms were ignored). In addition, COL records were only counted if they had "taxonRank" equal to "species" (i.e., subspecies were not counted) and "isExtinct" equal to "false."

iNaturalist classifies verifiable observations with photos, locations, and the date the observed species was identified to its species as either "needs ID" or "research grade," with "research grade" status requiring identification agreement among at least two-thirds of the identifiers (inaturalist.org/pages/help). Through the end of 2019, 68% of the records were identified to species, and 55% were classified as *research grade*. Although the iNaturalist platform shifted to include computer vision image recognition to aid in identification of species in 2017 (www.inaturalist.org/pages/computer_vision_demo), the proportion of observations that were research grade had not changed over time, with 55%–61% of observations reaching that quality threshold consistently over time (see supplemental figure S1).

All of our analyses were conducted in R version 3.6.1 (R Core Team 2019). All of the data used in this article are publicly available, and the code and data to replicate the analyses and reproduce the figures are available through GitHub (https://github.com/hurlbertlab/inat-user-behavior).

## Spatiotemporal and taxonomic patterns in iNaturalist observations

iNaturalist has been growing exponentially in both users and observations (figure 1a), with over 74 million total observations, 1.7 million observers, and 342,000 species documented as of July 2021 (www.inaturalist.org/observations). User activity is highest from May to September but with an increase in activity through the month of April and substantial spikes in total observations and unique users during organized events, such as the City Nature Challenge in late April, during which organizers in cities hold global events to encourage participants to record as many observations as they can in a single weekend (see https://citynaturechallenge.org; figure 1b). City Nature Challenges are a

specific instance of a wider effort by iNaturalist to encourage and provide infrastructure to support bioblitzes, which are communal efforts to record as many species in a given location and time period as possible (www.inaturalist.org/pages/bioblitz+guide). These uneven, intense concentrations of observer effort are a feature of iNaturalist and may be beneficial for some research questions (e.g., inventories of biodiversity) but detrimental to others (e.g., characterizing phenology).

iNaturalist activity varied substantially by the day of the week as well, with the total number of observations per day 37% higher on weekends than on weekdays, and the total number of user days (sum of unique users per day) was 22% higher on weekends (figure 1c). Globally, the observations were concentrated in North America, especially in the United States. Countries in South America, Europe, and Australia also have relatively high numbers of observations, although there are fewer observations in Western and Central Africa, Central America, and Southeast Asia (figure 1d). iNaturalist observations in the conterminous United States were disproportionately from developed areas and mixed and deciduous forests in which people live or might spend recreational time outdoors. Conversely, there were proportionally fewer observations in grasslands, shrublands, and agricultural areas, which may disproportionately be rural and privately owned and, therefore, difficult to access for citizen scientists in general (figure 1e). Notably, the overrepresentation of developed areas becomes even greater when examining only observations made by casual users (fewer than five observations total), with 58% of the observations from this group coming from developed areas compared with 38% of the observations by all users, although the true percentage of land area of the coterminous United States that is developed is 5% (see supplemental figure S2).

## Do iNaturalist users specialize taxonomically?

To examine taxonomic specialization of iNaturalist observers, we used hierarchical agglomerative clustering (HAC) with complete link (Lance and Williams 1967) to group users by the proportion of observations identified to species in different taxonomic groups. We performed two HAC analyses based on Euclidean distance matrices of each user's proportion of observations across taxonomic classes or insect orders. We clustered users with at least 50 observations identified to species on the basis of the proportion of their observations that fell into the top 10 classes on iNaturalist by the number of records: Agaricomycetes, Amphibia, Arachnida, Aves, Insecta, Liliopsida, Magnoliopsida, Mammalia, Polypodiopsida, and Reptilia. We retained the top 10 largest clusters identified by the HAC analysis, which was enough groups to capture variation in observer behavior while eliminating most small groups of only a few users. Second, we clustered users with at least 20 Insecta observations identified to species on the basis of the proportion of their observations that fell into the 25 insect orders that had at least 1000 total records in iNaturalist. We retained all clusters identified by the HAC analysis that included more than 10 users, resulting in a total of eight groups.

We quantified the degree of taxonomic specialization of the observers relative to a null expectation. For each user included in the clustering analysis, we calculated the Shannon evenness index ($E_{\text{observed}}$; Shannon 1948) on the basis of the proportion of observations in each taxonomic class or each insect order. We then calculated a $z$-score for each user's Shannon evenness index relative to a null distribution of possible Shannon evenness indices obtained by sampling 999 times from the list of all classes (218) or insect orders (27) weighted by the proportion of total iNaturalist observations per class or order. This weighting takes into account the fact that different taxonomic groups vary in their abundance and susceptibility to being documented. We also conducted analyses weighting by the total number of species per class or order as determined by the Catalog of Life with identical results. We calculated the $z$-transformed Shannon evenness index for each user using the mean ($E_{\text{null}}$) and standard deviation ($sd_{E\,\text{null}}$) of the null distribution of evenness values as follows:

$$E_{z-\text{transformed}} = \frac{E_{\text{observed}} - \overline{E}_{\text{null}}}{sd_{E\,\text{null}}}$$

Negative values indicate users that are more specialized taxonomically, whereas values close to 0 indicate users that sample taxonomic classes or orders roughly in proportion to their density or diversity. Strong positive values were uncommon but would indicate users that have a much more even representation of observations across classes or orders than would be expected by chance. We conducted a sensitivity analysis of our user activity and taxonomic specialization analyses including only research grade observations and found that the results were qualitatively very similar to results including observations needing identification and casual observations as well (see supplemental table S1, figures S3–S5), so in the present article, we present results based on all observations identified to species.

Among the users with multiple submissions to iNaturalist, the hierarchical clustering results reveal how the observations were distributed across classes. Of the users with at least 50 identified observations, 51% focused primarily on plants (Liliopsida and Magnoliopsida) and insects, the most common and diverse groups of terrestrial organisms that are difficult to miss because they are numerous, visible, and photographable (figure 2a). The users in the second most common group (about 30% of users) focused almost exclusively on plants, whereas the users in the third most common group (15% of users) had a strong focus on birds (figure 2a). The remaining user groups were defined by specializations on Agaricomycetes (which includes mushroom-forming fungi; less than 2%), ray-finned fishes, reptiles, amphibians, mammals, monocots (Liliopsida), or arachnids (each group less than 1%; figure 2a). The users in all

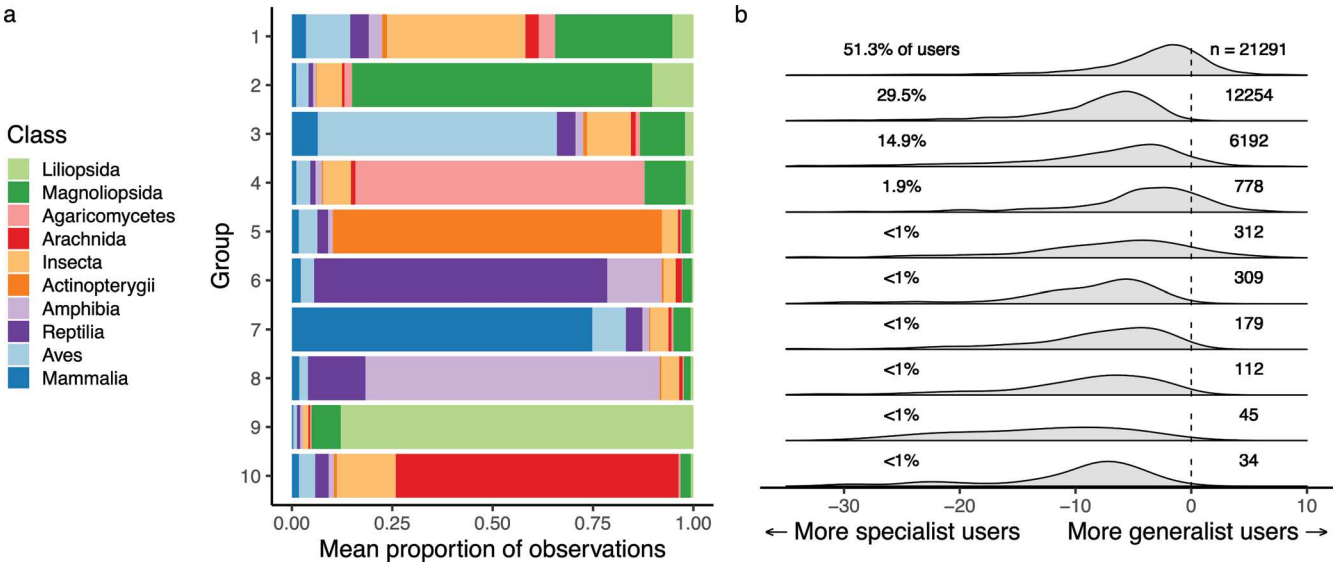


***Figure 2. Taxonomic specialization of iNaturalist users with at least 50 observations through 31 December 2019 at the class level showing the composition of observations in each group of observers on the left and the distribution of specialization z-scores for users in that group on the right, compared to a null expectation of evenly observing species across classes. Users were grouped using agglomerative hierarchical clustering on the basis of the proportion of observations by class for the top 10 classes in iNaturalist by total number of records, and proportion of observations from each class are shaded by color. (a) Barplot showing the average proportion of observations per class for users in each group. (b) Density plots for each group showing the relative specialization of users in that group. Negative values indicate higher specialization across classes than expected, values of 0 (vertical dashed line) indicate that classes are represented as expected on the basis of their overall prevalence within the iNaturalist data set, and positive values indicate observations distributed more evenly across classes than expected. To the left of each distribution is the percentage of users that fall into that group, and to the right is the number of users in each group.***

of these groups tended to be more specialized taxonomically than expected from the null (figure 2b); however, the plant–insect group (group 1) and the Agaricomycetes-biased group (group 4) each included a substantial fraction of users that submitted observations of different classes. Across all of the groups, 77% of the users had specialization indices less than –1.96, indicating specialization at the class level much greater than the null expectation.

We also examined the degree of specialization on particular orders within Insecta. Most users with at least 20 insect observations (about 60%) could be considered insect generalists, with the distribution of evenness indices centered near zero, although this generalist group (group 1) had a plurality of observations in the order Lepidoptera (figure 3a and 3b). Showy, charismatic, and conspicuous groups, including Lepidoptera, Odonata, and Hymenoptera, were the most common groups to specialize on, with 34% of the users focused almost exclusively on Lepidoptera or Lepidoptera and Odonata. Three percent of the users fell into a group with a majority of observations in Hymenoptera (figure 3a), possibly representing users focused specifically on pollinators. A handful of users with the most extreme departures from the null expectation of evenness were highly specialized on Hemiptera, Diptera, and Orthoptera (figure 3a and 3b). Across all Insecta groups, 32% of the users had specialization indices less than –1.96, indicating that specialization on particular orders within Insecta was less common than specialization by users at the class level. Future work could examine whether this switch from specializing observations at broad taxonomic scales and generalizing at more narrow taxonomic scales holds true in other groups besides insects.

The completeness of taxonomic coverage in iNaturalist observations varied substantially with respect to taxonomic class. Many plant, animal, and fungi classes had fewer than 25% of known, extant species recorded and identified in iNaturalist, including some of the most species rich, such as Insecta (figure 4a). Classes with the most complete record of species in iNaturalist either have very few species per class (e.g., Ginkoopsida, ginkgo trees; Merostomata, horseshoe crabs) or are highly visible and interesting to humans and may be easier to photograph (e.g., Pinopsida, Reptilia, Mammalia). In particular, nearly 90% of extant bird species have at least one observation recorded in iNaturalist. The most commonly observed species on iNaturalist tended to be species that are easily observed by virtue of their size, behavior, and conspicuousness and that are common in human-influenced areas, including monarch butterflies

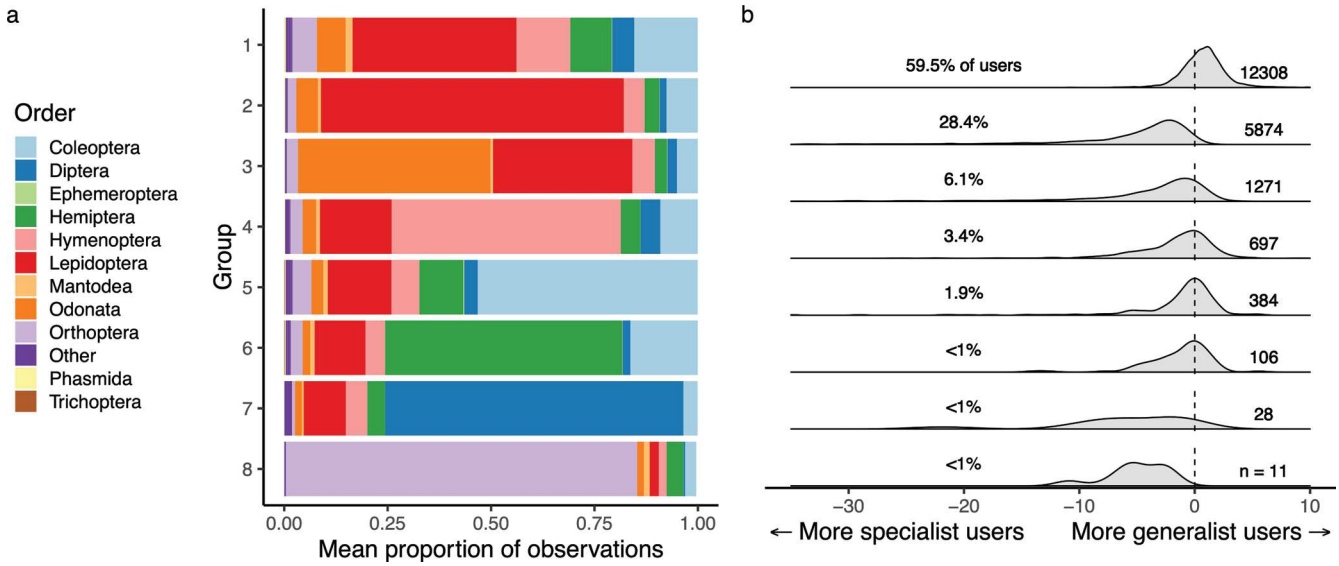


***Figure 3. Taxonomic specialization of iNaturalist users with at least 20 Insecta observations through 31 December 2019 by Insecta order, showing the composition of observations in each group of observers on the left and the distribution of specialization z-scores for users in that group on the right, compared to a null expectation of evenly observing species across orders. The users were grouped using agglomerative hierarchical clustering on the basis of the proportion of observations by order for orders in Insecta with at least 1000 records in iNaturalist. The shaded colors show the 11 Insecta orders with the highest total number of records, and the "other" category is the sum of observations made of all other orders. (a) Barplot showing the average proportion of observations per order by proportion of observations for users in each group. (b) Density plots for each group showing the relative specialization of users in that group. Negative values indicate higher specialization across orders than expected, values of 0 (vertical dashed line) indicate that classes are represented as expected on the basis of their overall prevalence within the iNaturalist data set, and positive values indicate observations distributed more evenly across orders than expected. To the left of each distribution is the percentage of users that fall into that group.***

(*Danaus plexippus*), mallards (*Anas platyrhynchos*), and eastern gray squirrels (*Sciurus carolinensis*; table 1). The 10 most observed species on iNaturalist represent 4% of the observations of the casual users (fewer than five observations all time) and 3% of the total observations of the users with five or more observations over time. The taxonomic groups that were most well documented tended to be easy to photograph, abundant (e.g., plants and insects), or especially interesting to humans (e.g., butterflies and birds).

### Variation in user activity levels

We characterized annual user frequency and intensity during the most active months of iNaturalist usage each year, May to September. For the users that submitted at least one observation, we calculated the median number of observations per day, the number of observation dates per year (omitting the first year a user was active to exclude cases in which a user joined the platform midseason), and the total number of observations per user during this period. Most users on iNaturalist are infrequent even during the most active months on the platform by total observations, with 50% of the users active three or fewer days per year, making three or fewer observations, and uploading not more than one observation per active day from May through September (table 2). However, there were a small number of very active users, with the top 5% of the users based on the number of active days per year active at least 37 days per year and the top 5% of the users by observations per day making more than 15 observations per day (table 2). The observed pattern that a large majority of observations identified to species come from only a handful of observers has been observed in other citizen science projects (August et al. 2020) and more generally in economics and other fields (known as the Pareto principle; Newman 2005).

To evaluate the tendency of users to submit repeat observations of species they had previously recorded, we examined the relationship between the total number of observations and the number of species observed. Users that primarily use iNaturalist as a means of collecting new species to maximize a personal species list or who use iNaturalist to help identify species they have not seen before will fall close to the one-to-one line on such a plot, whereas users who submit repeated observations of species will deviate from the one-to-one line with many more observations than species. The majority of iNaturalist users could indeed be characterized as low-frequency collectors.

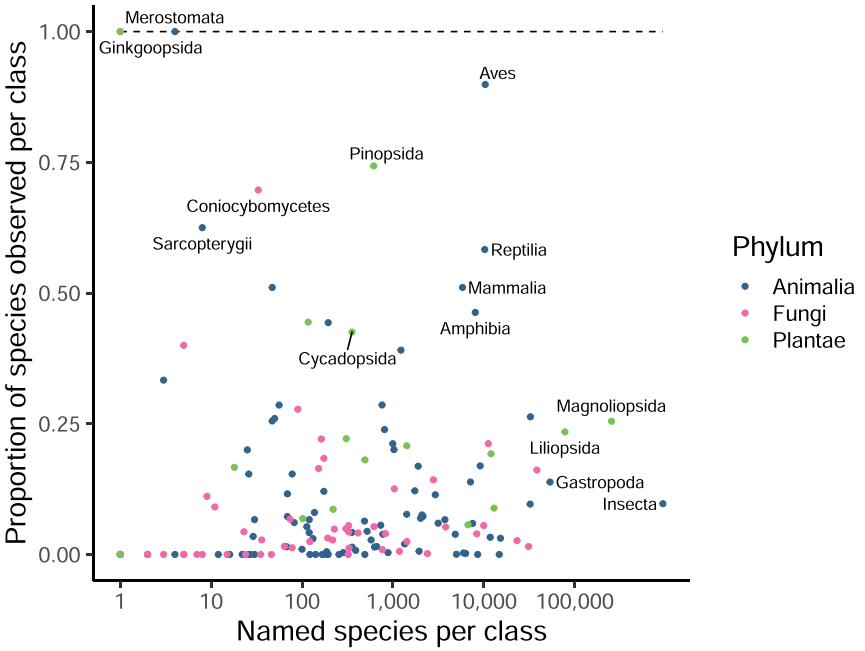


***Figure 4. Proportion of extant species per class with at least one record in iNaturalist through 31 December 2019 compared with the known number of extant species per class from the Catalogue of Life. The color of the dots indicates the phylum.***

When including just iNaturalist observations identified to species, the median user made six observations of six different species, with the most common case being users making only one observation of one species (figure 5a). Even users with dozens to hundreds of observations tended to make close to one observation per species. The most active users, defined in the present article as users with more than 1000 observations, begin to depart from that one-to-one relationship between recorded species and observations (figure 5a). Notably, it is these highly active users that contributed most of the observations that had been identified to species. The top 10% of the users provided about 87% of the observations identified to species, whereas the top 1% (users with more than 455 observations identified to species) contributed about 62% of identified observations (figure 5b). More active users on the platform may be more involved in organized activities that encourage repeated posting of observations, such as City Nature Challenge, bioblitzes, and other organized events, which could contribute to this observed pattern. The more active users had typically been contributing observations for multiple years, and even those who had submitted only a single observation per year (e.g., first observation of the season) were more likely to accumulate repeat observations of a species across years.

## Implications of the iNaturalist observation process for biodiversity research

As we have shown, iNaturalist observations can be biased taxonomically in several ways, many of which are common across opportunistic records in general (Isaac and Pocock 2015). Although there are certain organisms that are easier to photograph with widely available smartphone cameras (although many submissions include photographs taken with traditional cameras or other methods of recording a species, such as an audio recording or spectrogram), organisms will be undersampled if they are often hidden from view, highly mobile, not able to be identified on the basis of photographs alone, or likely to avoid close approach by humans. These categories might include many larger animals and organisms that are not often found out in the open (Hochmair et al. 2020), as well as insects such as flies that are less likely to remain still long enough to be photographed. Organisms of very small size are difficult to photograph clearly without special equipment and are more difficult to identify to species (Unger et al. 2020). For some organisms, including many insects, proper identification to species requires clear views of genitalia or other body structures that may not be evident in photographs, and in some cases, identification may only be possible on dissection or sequencing.

The observations from iNaturalist are widespread and numerous for many taxonomic groups, but appropriate use of these data in research applications requires careful consideration of the sampling process and user behavior on the platform. The observations are well suited to research questions that can be answered with nonuniformly sampled presence-only observations, including cataloging species lists of organisms that are likely to be observed and identified by the iNaturalist community in ecosystems with high observation density or tracking species invasions (Prudic et al. 2018, Hiller and Haelewaters 2019, Leong and Trautwein 2019).

In addition to the limitations associated with presence-only observations in general (Dorazio 2012, Yackulic et al. 2013), iNaturalist observations may be biased in directions that diverge from other common sources of presence-only data. For example, although iNaturalist observations have a strong bias toward developed areas, specimen records from museum collections are becoming less biased toward human-influenced areas over time (Shirey et al. 2021). In particular, the spatial bias in iNaturalist records may be important to consider in applying these data in species distribution modeling or measuring habitat associations of species, because developed areas will be overrepresented in the records, whereas harder to access, more remote areas and habitats will be underrepresented. This pattern is exacerbated in observations made by infrequent or casual users of the platform, and restricting analyses to more

***Table 1. Top ten most observed species in iNaturalist and the total number of observations of those species through 31 December 2019.***

| Common name | Scientific name | Order | Number of observations |
|---|---|---|---|
| Monarch | *Danaus plexippus* | Lepidoptera | 73,929 |
| Western honey bee | *Apis mellifera* | Hymenoptera | 70,473 |
| Mallard | *Anas platyrhynchos* | Anseriformes | 69,916 |
| Great blue heron | *Ardea herodias* | Pelecaniformes | 53,124 |
| Canada goose | *Branta canadensis* | Anseriformes | 45,756 |
| Red-tailed hawk | *Buteo jamaicensis* | Accipitriformes | 45,681 |
| House sparrow | *Passer domesticus* | Passeriformes | 45,289 |
| Great egret | *Ardea alba* | Pelecaniformes | 45,122 |
| American robin | *Turdus migratorius* | Passeriformes | 44,250 |
| Eastern gray squirrel | *Sciurus carolinensis* | Rodentia | 44,157 |

***Table 2. iNaturalist user observation frequency during the summer months, May through September, including all observations made through 30 September 2019.***

| Metric | Median | 5th percentile | 95th percentile |
|---|---|---|---|
| Active days per year | 3 | 1 | 37 |
| Observations per day | 1 | 1 | 15 |
| Observations per user | 3 | 1 | 64 |

*Note:* When calculating the number of dates per year for each user, we included only the second year of activity for a given user to avoid cases in which a user joined the platform midyear.

active users may reduce these spatial biases by yielding a more representative distribution of sampled habitats, improving species distribution models built with these data (Van Eupen et al. 2021). Because downloaded iNaturalist records are associated with a username, understanding these spatial biases in a specific subset of the data is readily achievable, and the methods that have been developed for other opportunistic citizen science projects (e.g., Kelling et al. 2019, August et al. 2020) may be useful to researchers using iNaturalist data as well. Other promising methods include incorporating information about species observed by users other than the target species, which has been shown to improve species distribution model performance (Milanesi et al. 2020).

iNaturalist data clearly contain a signal of phenology of individual species, but characterizing phenology accurately requires considering how temporal variation in observer effort might obscure, bias, or exaggerate the underlying pattern. The weekend effect that we documented in iNaturalist and that has been observed in other citizen science data sets (Courter et al. 2013) may pose a problem specifically for inferring how phenology has shifted over time. This is because the day of the year for a given day of the week (e.g., the first Saturday in June) advances 1 day earlier each year (and 2 days in a leap year) until resetting roughly every 7 years. Therefore, an examination of phenological timing over a time series of just 5–10 years might suggest a shift toward earlier phenology—even when there was no true underlying trend—because of the shift in the timing of weekend observations. Aggregating observations to a temporal resolution of 1 week would eliminate this bias, although the reduced temporal resolution may introduce uncertainty to phenometrics in data sets with a small total number of observations. Certain phenometrics may also be biased by the continued exponential growth of the iNaturalist platform over time. As the number of observers and observations increases each year, first or last observations dates, in particular, but even quantiles such as the 5th and 95th dates could become biased toward more extreme values (Belitz et al. 2020, Park et al. 2021). For particular species, the ability to estimate phenology will depend on how easily the organism is observed, photographed, and identified. In addition, users may differ in whether they are more likely to record only the first of a species they observe in a given year or to make repeated observations.

As the temporal extent of iNaturalist data grows, observations may increasingly be used to estimate trends in biodiversity over time (such as tracking occurrence over time), although careful consideration must be given to biases in the iNaturalist data set in these applications including increasing numbers of observations over time and seasonality of observations within years. Qualitatively similar patterns have been found in population trends estimates from standardized surveys and iNaturalist observations (controlling for effort by standardizing using the total number of iNaturalist observations) in butterfly species in western North America (Forister et al. 2021). Standardizing observations of a particular species using general patterns of iNaturalist activity may be effective at removing bias associated with individual events such as City Nature Challenges but may prove challenging over longer periods of time, especially when biological patterns in the species of interest mirror patterns in the greater iNaturalist data set (e.g., spring insect emergence concurrent with a seasonal increase in observations on the platform). Another recent effort to capture biodiversity change using iNaturalist records made use of record metadata to reconstruct observation events and species lists for observers and showed positive correlations between rank change in California coastal species when compared with estimates from standardized surveys (Rappaciuolo et al. 2021). Furthermore, comparisons of trends between species with different likelihoods of detection

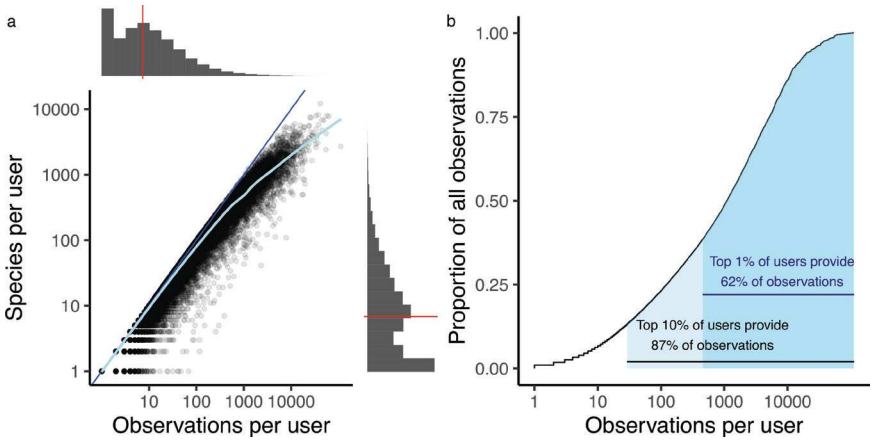


***Figure 5. iNaturalist user activity for observations identified to species through 31 December 2019. (a) The number of species observed by a user compared to total number of observations. The blue line represents a one-to-one relationship, and the light blue line shows a smoothed generalized additive model fit. The marginal histograms show the distribution of species per user and observations per user, with the red lines showing the median of those distributions. (b) The cumulative proportion of all observations identified to species by user's total number of observations. Shaded regions show the 90th and 99th percentiles of users by total number of observations.***

are complicated by the lack of detailed survey effort. Even comparing relative, instead of absolute, abundance between species may be difficult using these data without accounting for differences in ease of detection and documentation.

In our analyses characterizing user taxonomic specialization, we used observations identified to species. Observations are often identified by participants on the platform different than the user uploading the observation, and for the observation to attain research grade status, multiple users must agree. As a result, the identification process itself is an important component of the iNaturalist platform and community, although it is beyond the scope of this article. Nevertheless, important areas for future inquiry include a detailed exploration of the behavior and activity level of identifiers, the resolution of identification disagreements, and how the time to identification (18 days on average; www.inaturalist.org/stats) varies by taxonomic group or geographic region.

Our results are a broad first look at how users on iNaturalist use the platform and can provide a starting point for considering what portion of iNaturalist users or observations may be most relevant to a particular question. Researchers might prefer to use observations from iNaturalist users on the higher end of the activity spectrum, who are more likely to record more than one observation of a species or exhaustively sample a particular taxonomic group of interest in their local area. Because a vast majority of higher-quality observations come from these very active users, excluding low-activity users (casual or one-off accounts) will not be a large penalty on the sample size of observations available to use. Despite featuring biases associated with opportunistic, presence-only observations, the vast number of engaged observers and identifiers and extent of observations on iNaturalist, combined with associated photographs of the organisms, makes the project and data generated from it an invaluable resource to biodiversity researchers.

## Conclusions

Our results build on previous work that demonstrates the importance of examining biases and filtering approaches on a case-by-case basis when working with massive citizen science data (Steen et al. 2019), and researchers should consider whether their analysis is affected by systematic biases in observations from one-off, casual users or highly active users. We suggest that simple corrections for observer behavior such as normalizing observations by the total observations in a given time period may be insufficient if the research interest is in monitoring distribution or abundance changes. Further modeling of the observation process to understand how user behavior may bias biodiversity estimates will be essential in developing toolkits for leveraging unstructured citizen science observations to address questions in biodiversity science. iNaturalist as a platform provides substantial value not only as a tool for researchers but as a place for community building and connecting with other naturalists as well.

## Acknowledgments

Support was provided by funding from National Science Foundation grants no. EF-1702708 and no. EF-1703048. MB was supported by a University of Florida Biodiversity Institute fellowship. We would like to thank Carrie Seltzer and Ken-ichi Ueda for support and review of initial drafts. We are especially indebted to the thousands of observers and identifiers on the iNaturalist platform, without whom this work would not have been possible.

## Supplemental material

Supplemental data are available at *BIOSCI* online.

## References cited

August, T, Fox, R, Roy, DB, Pocock, MJO. 2020. Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. Scientific Reports 10: 11009. https://doi.org/10.1038/s41598-020-67658-3.

Balaguera-Reina SA, Bustillo S, Zarrate-Charry DA, Charry F, Cepeda-Mercado AA, González-Maya JF. 2019. Conservation status and distribution based on a species distribution model of the endemic yellow-striped poison frog, *Dendrobates truncatus* (Cope, 1861), in Colombia. Herpetological Review 50: 52–57.

Barve VV, et al. 2020. Methods for broad-scale plant phenology assessments using citizen scientists' photographs. Applications in Plant Sciences 8: e11315.

Belitz MW, Larsen EA, Ries L, Guralnick RP. 2020. The accuracy of phenology estimators for use with sparsely sampled presence-only observations. Methods in Ecology and Evolution 11: 1273–1285.

Boakes EH, Gliozzo G, Seymour V, Harvey M, Smith C, Roy DB, Haklay M. 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. Scientific Reports 6: 33051. https://doi.org/10.1038/srep33051.

Bonney R, Shirk JL, Phillips TB, Wiggins A, Ballard HL, Miller-Rushing AJ, Parrish JK. 2014. Next steps for citizen science. Science 343: 1436–1437. https://doi.org/10.1126/science.1251554.

Brown ED, Williams BK. 2019. The potential for citizen science to produce reliable and useful information in ecology. Conservation Biology 33: 561–569.

Callaghan CT, Ozeroff I, Hitchcock C, Chandler M. 2020. Capitalizing on opportunistic citizen science data to monitor urban biodiversity: A multi-taxa framework. Biological Conservation 251: 108753. https://doi.org/10.1016/j.biocon.2020.108753.

Chapman D, Prescott OL, Roy HE, Tanner R. 2019. Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. Journal of Biogeography 46: 1029–1040.

Chardon NI, Cornwell WK, Flint LE, Flint AL, Ackerly DD. 2015. Topographic, latitudinal and climating distribution of *Pinus coulteri*: Geographic range limits are not at the edge of the climate envelope. Ecography 38: 590–601.

Courter JR, Johnson RJ, Stuyck CM, Lang BA, Kaiser EW. 2013. Weekend bias in citizen science data reporting: Implications for phenology studies. International Journal of Biometeorology 57: 715–720.

Dorazio RM. 2012. Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68: 1303–1312.

Drury JP, Barnes M, Finneran AE, Harris M, Grether GF. 2019. Continent-scale phenotype mapping using citizen scientists' photographs. Ecography 42: 1436–1445.

Follett R, Strezov V. 2015. An analysis of citizen science based research: Usage and publication patterns. PLOS ONE 10: 143687.

Forister ML, et al. 2021. Fewer butterflies seen by community scientists across the warming and drying landscapes of the American West. Science 371: 1042–1045.

Fourcade Y. 2016. Comparing species distributions modelled from occurrence data and from expert-based range maps: Implication for predicting range shifts with climate change. Ecological Informatics 36: 8–14.

Gazdic M, Groom Q. 2019. iNaturalist is an unexplored source of plant–insect interaction data. Biodiversity Information Science and Standards 3: e37303.

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021. Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118: e2018093118.

Hiller T, Haelewaters D. 2019. A case of silent invasion: Citizen science confirms the presence of *Harmonia axyridis* (Coleoptera, Coccinellidae) in Central America. PLOS ONE 14: 220082.

Hochmair HH, Scheffrahn RH, Basille M, Boone M. 2020. Evaluating the data quality of iNaturalist termite records. PLOS ONE 15: 226534.

Isaac NJ. Pocock MJO. 2015. Bias and information in biological records. Biological Journal of the Linnean Society 115: 522–531.

Kelling S, et al. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. BioScience 69: 170–179.

Lance GN, Williams WT. 1967. A general theory of classificatory sorting strategies: 1. Hierarchical systems. Computer Journal 9: 373–380.

Leong M, Trautwein M. 2019. A citizen science approach to evaluating US cities for biotic homogenization. PeerJ 7: e6879. https://doi.org/10.7717/peerj.6879.

Li D, et al. 2020. Climate, urbanization, and species traits interactively drive flowering duration. Global Change Biology 27: 892–903.

Milanesi P, Mori E, Menchetti M. 2020. Observer-oriented approach improves species distribution models from citizen science data. Ecology and Evolution 10: 12104–12114.

Moulin N. 2020. When citizen science highlights alien invasive species in France: The case of Indochina mantis, *Hierodula patellifera* (Insecta, Mantodea, Mantidae). Biodiversity Data Journal 8: e46989.

Newman MEJ. 2005. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46: 323–351.

Park DS, Newman EA, Breckheimer IK. 2021. Scale gaps in landscape phenology: Challenges and opportunities. Trends in Ecology and Evolution 36: 709–721.

Pocock MJO, Tweddle JC, Savage J, Robinson LD, Roy HE. 2017. The diversity and evolution of ecological and environmental citizen science. PLOS ONE 12: 172579.

Prudic KL, Oliver JC, Brown BV, Long EC. 2018. Comparisons of citizen science data-gathering approaches to evaluate urban butterfly diversity. Insects 9: 186. https://doi.org/10.3390/insects9040186.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rappaciuolo G, Young A, Johnson R. 2021. Deriving indicators of biodiversity change from unstructured community-contributed data. Oikos 130: 1225–1239. https://doi.org/10.1111/oik.08215.

Roskov Y, et al., eds. 2020. Species 2000 and ITIS Catalogue of Life. Naturalis.

Shannon C. 1948. A mathematical theory of communication. Bell System Technical Journal 27: 623–656.

Shirey V, Belitz MW, Barve V, Guralnick R. 2021. A complete inventory of North American butterfly occurrence data: Narrowing data gaps, but increasing bias. Ecography 44: 537–547.

Steen VA, Elphick CS, Tingley MW. 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. Diversity and Distributions 25: 1857–1869.

Unger S, Rollins M, Tietz A, Dumais H. 2020. iNaturalist as an engaging tool for identifying organisms in outdoor activities. Journal of Biological Education 2020: 1739114. https://doi.org/10.1080/00219266.2020.1739114.

Van Eupen C, Maes D, Herremans M, Swinnen KRR, 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.

Welvaert M, Caley P. 2016. Citizen surveillance for environmental monitoring: Combining the efforts of citizen science and crowdsourcing in a quantitative data framework. SpringerPlus 5: 1890. https://doi.org/10.1186/s40064-016-3583-5.

Werenkraut V, Baudino F, Roy HE. 2020. Citizen science reveals the distribution of the invasive harlequin ladybird (*Harmonia axyridis* Pallas) in Argentina. Biological Invasions 22: 2915–2921.

Wilson JS, Pan AD, General DEM, Koch JB. 2020. More eyes on the prize: An observation of a very rare, threatened species of Philippine bumble bee, *Bombus irisanensis*, on iNatuarlist and the importance of citizen science in conservation biology. Journal of Insect Conservation 24: 727–729.

Yackulic CB, Chandler R, Zipkin EF, Royle JA, Nichols JD, Grant EHC, Veran S. 2013. Presence-only modelling using MAXENT: When can we trust the inferences? Methods in Ecology and Evolution 4: 236–243.

*Grace Di Cecco (gdicecco@live.unc.edu) is a PhD student and Allen Hurlbert is a professor in the Department of Biology at the University of North Carolina, in Chapel Hill, North Carolina, in the United States. Michael Belitz is a PhD student, Vijay Barve is a postdoctoral research associate, Brian Stucky is an assistant scientist, and Robert Guralnick is curator of biodiversity informatics at the Florida Museum of Natural History, in Gainesville, Florida, in the United States.*