Ornithological Applications

AmericanOrnithology.org

Volume 123, 2021, pp. 1–17
https://doi.org/10.1093/ornithapp/duab038

*RESEARCH ARTICLE*

# Distribution models using semi-structured community science data outperform unstructured-data models for a data-poor species, the Plain Tyrannulet

**Fabricio C. Gorleri,[1,2] Wesley M. Hochachka,[3] and Juan I. Areta[1]**

[1] Laboratorio de Ecología, Comportamiento y Sonidos Naturales, Instituto de Bio y Geociencias del Noroeste Argentino (IBIGEO-CONICET), Salta, Argentina
[2] Science Department, Aves Argentinas/Asociación Ornitológica del Plata, Buenos Aires, Argentina
[3] Cornell Laboratory of Ornithology, Ithaca, New York, USA
*Corresponding author: fabriciogorleri@gmail.com

Submission Date: April 25, 2021; Editorial Acceptance Date: July 16, 2021; Published August 28, 2021

## ABSTRACT

Modeling the distribution of a data-poor species is challenging due to a reliance on unstructured data that often lacks relevant information on sampling and produces coarse-resolution outputs of varying accuracy. Data on sampling effort associated with higher-quality, semi-structured data derived from some community science programs can be used to produce more precise models of distribution, albeit at a cost of using fewer data. Here, we used semi-structured data to model the seasonal ranges of the Plain Tyrannulet (*Inezia inornata*), a poorly known Austral–Neotropical migrant, and compared predictive performance to models built with the full unstructured dataset of the species. By comparing these models, we examined the relatively unexplored tradeoff between data quality and data quantity for modeling of a data-sparse species. We found that models using semi-structured data outperformed unstructured-data models in the predictive accuracy metrics (mean squared error, area under the curve, kappa, sensitivity, and specificity), despite using only 30% of the available detection records. Moreover, semi-structured models were more biologically accurate, indicating that the tyrannulet favors arboreal habitats in dry and hot lowlands during the breeding season (Chaco region) and is associated with proximity to rivers in tropical and wet areas during the nonbreeding season (Pantanal, Beni, and southwest Amazonia). We demonstrate that more detailed insights into distributional patterns can be gained from even small quantities of data when the data are analyzed appropriately. The use of semi-structured data promises to be of wide applicability even for data-poor bird species, helping refine information on distribution and habitat use, needed for effective assessments of conservation status.

*Keywords:* citizen science, data limitations, data scarcity, eBird, *Inezia inornata*, migration, Neotropics

## LAY SUMMARY

- Modeling the distributions of poorly known species is compromised by the sparse and noisy available data, often leading to coarse-resolution models.
- Semi-structured community science data, capable of accounting for sources of biases, can provide more accurate insights into species' distributions, but their effectiveness remains unclear with small datasets.
- We evaluated the performance of models built with semi-structured data for a data-poor species (Plain Tyrannulet) against models built with all the available records of the tyrannulet that maximize sample size but for which variation in the sampling process could not be corrected.
- The predictive accuracy of models was better when using semi-structured data, even at the expense of a 70% to 72% reduction in the number of detection records.
- We demonstrated that improved information on distribution and habitat use can result from even small quantities of high-quality data, information that is critical for an effective conservation assessment of currently poorly known species.

**Modelos de distribución con datos semiestructurados de ciencia comunitaria superan a modelos con datos no estructurados para una especie con escasos datos, el *Inezia inornata***

## RESUMEN

El modelado de distribuciones de especies con pocos datos es complejo, ya que suelen utilizarse datos no estructurados que carecen de metadatos relevantes de esfuerzo de muestreo, resultando en modelos de baja resolución. Los metadatos de esfuerzo de muestreo contenidos en datos semiestructurados de "mayor calidad" de ciencia ciudadana

pueden utilizarse para producir modelos de distribución más precisos, aunque a costas de utilizar una menor cantidad de datos. Aquí utilizamos datos semiestructurados para modelar los rangos estacionales del Piojito Picudo (*Inezia inornata*) —un migrante Austral poco conocido del Neotrópico—, y comparamos el rendimiento predictivo frente a modelos construidos con el conjunto de datos completo de la especie. Al compararlos, examinamos el compromiso poco explorado que existe entre cantidad y calidad de datos para modelar la distribución de una especie con pocos datos. Encontramos que los modelos con datos semiestructurados superaron a los modelos de datos no estructurados en las métricas de precisión (MSE, AUC, Kappa, Sensibilidad, y Especificidad), a pesar de haber utilizado sólo el 30% de los registros de detecciones disponibles. Además, los modelos con datos semiestructurados fueron biológicamente más precisos: identificaron que el piojito utiliza hábitats arbóreos en tierras bajas calurosas y secas durante la temporada de cría (región del Chaco), pero que está asociado a la proximidad de ríos en regiones tropicales y húmedas durante el invierno (Pantanal, Beni y SO de la Amazonia). Demostramos que es posible obtener una visión más detallada de patrones distribucionales incluso con poca cantidad de datos cuando son analizados adecuadamente. El uso de datos semiestructurados promete ser de amplia aplicación incluso para especies de aves con pocos datos, ayudando a mejorar la información disponible sobre la distribución y el uso del hábitat, información necesaria para evaluar eficazmente el estado de conservación.

*Palabras clave:* ciencia ciudadana, datos escasos, datos limitados, *Inezia inornata*, migración, Neotrópico

## INTRODUCTION

The emergence of digital databases of species occurrences, largely powered by community science initiatives (also referred to as "citizen science"), opens new doors to increase our understanding of distributional patterns of organism (Hampton et al. 2013, Feldman et al. 2021). This is particularly useful for those species that were historically poorly studied, for which the information about their basic ecology is still coarse-grained or even lacking. While these new data sources provide an opportunity to describe distributions at resolutions that are more relevant to conservation (Sumner et al. 2019), it is still challenging to generate accurate distributional information when data are sparse or when most data are noisy and biased given the haphazard schemes of data collection (Brotons et al. 2004). In such cases, deciding which data to use to accurately describe the distribution of data-poor species is a difficult task. However, the resulting outputs are critical to correctly uncover key environmental drivers of distributions and to inform conservation planning (Elith and Leathwick 2009).

Species distribution models are a popular tool that uses georeferenced records to predict species ranges and species–habitat relationships at fine spatial resolutions (Guisan and Zimmerman 2000). Ideally, models should be built with large and high-quality datasets (Brotons et al. 2004). Larger datasets provide a greater number of observations for model training and usually result in broader geographic and environmental coverage of the samples, desirable to obtain accurate model outputs. Models also benefit from "high-quality" data containing survey effort metadata capable of accounting for sources of variation in the data collection process (Brotons et al. 2004, Johnston et al. 2021). Information about the effort expended during each survey event is critical to account for variation in detection rates in relation to sampling effort. As such, models incorporating survey effort metadata may better explain whether the lack of observations at a site reflects the absence of a species and estimate how much effort is needed to detect the species where present. Unfortunately, this "ideal modeling scenario" of having large and high-quality datasets is rarely achieved for data-poor species.

One common practice for modeling the distribution of a data-poor species is to pool all available records from diverse sources to be analyzed under the same analytical framework (Biddle et al. 2021). Because of the disparate nature of these records, most of the compiled data inevitably have a presence-only format and are unstructured (i.e. the locality and date are known, but no sampling effort metadata exist; La Sorte et al. 2018). Examples of these data are largely from museum records, biodiversity inventories, web reports, and presence-only community science programs or stored in Global Biodiversity Information Facility (GBIF, www.gbif.org), which is a worldwide repository of unstructured biodiversity data. Given the difficulties in obtaining information on species' absences with these unstructured data, modeling methods typically associate presence records with randomly generated background points that serve as imperfect surrogates of absence locations (presence-background methods; see Breiner et al. 2015). While presence-background methods are effective at maximizing the number of samples, the resulting models may be limited in their ability to produce robust ecological inferences as a consequence of the few informative data that are used (Guillera-Arroita et al. 2015).

In recent years, crowdsourcing efforts that rely on community science programs are rapidly producing more informative, "higher-quality" data at broad scales. These semi-structured data (Welvaert and Caley 2016, Kelling et al. 2019) not only are opportunistically collected by volunteers but also have metadata describing the observation process. These metadata can be used to address many of the problems arising with unstructured

data, in addition to inferring the absence locations of the species of interest (La Sorte et al. 2018, Kelling et al. 2019). Some community science initiatives such as eBird (Sullivan et al. 2009), eButterfly (Prudic et al. 2017), and iSeeMammals (Sun et al. 2018, preprint) collect semi-structured data through the design of specific but flexible sampling protocols that record information of observation start time, distance traveled, duration, number of observers, and whether all species detected were reported in a checklist. As these projects have gained popularity, semi-structured data are exponentially increasing across wider geographic areas (Callaghan et al. 2019, La Sorte and Somveille 2020) and are becoming more available in many regions for which only unstructured data were available in the past.

The desired effect of using semi-structured data for distribution modeling is to control sampling biases when creating the models (Johnston et al. 2021), but its effectiveness with data-poor species remains poorly assessed. Robust modeled products result from semi-structured data when datasets are rigorously analyzed and subjected to stringent data filtering to retain only the most informative records (Steen et al., 2019, Johnston et al. 2021). However, the number of samples is inevitably reduced by selecting only a subset of all the available data. This becomes problematic if the goal is to use semi-structured data to model the distributions of data-poor species, as any subsetting process on an already small dataset can result in a substantial loss of information. Of particular concern is that any filtering of the available data may reduce the geographic coverage of the detection data in addition to decreasing the quantity of the data. If that sample reduction also reduces the range of environmental conditions (e.g., habitat types or climate regimes) for which data are available, then the resulting models can be environmentally biased (Reese et al. 2005), producing unreliable outputs.

Although some studies have examined this common tradeoff between data quantity and quality (Steen et al. 2019, Van Eupen et al. 2021), its effects on data-poor species using recently available semi-structured data remain unexplored. Specifically, we believe that it is important to better understand for data-poor species whether the use of only a higher-quality set of records will be beneficial for modeling at the expense of reducing the number of samples or whether the larger sample sizes available from retaining all records will provide more information despite sampling noise and bias. The resulting information will be critical to advance our knowledge on how to better describe distributional ranges and habitat relationships of any species for which current information is coarse or nonexistent.

Bird species comprising the Austral–Neotropical migration system (i.e. migrating within the Neotropics) provide good candidates to evaluate the modeling performance of recently available semi-structured data. More than 230 bird species are Austral–Neotropical migrants (Chesser 1994, Stotz et al. 1996), but their breeding and nonbreeding grounds are still poorly understood (Jahn et al. 2020), and refined knowledge of their current spatiotemporal distributions is urgently needed. Recent studies using data derived from online databases to explore the migration ecology of these species have performed simple analyses within a presence-background framework with unstructured data (e.g., Areta and Bodrati 2010, Lees and Martin 2014, Lees 2016, Hayes et al. 2018, DeGroote et al. 2020, Biddle et al. 2021, Da Silveira et al. 2021). While informative, the spatial resolution of the outputs is coarse and partially biased toward more frequently sampled sites.

In this study, we used semi-structured data obtained from eBird to model the seasonal ranges of the Plain Tyrannulet (*Inezia inornata*, Aves: Tyrannidae), a poorly known Austral–Neotropical migrant, using Random Forest models. We compared the predictive performance of models from semi-structured data to other Random Forest models built using the full dataset of the species occurrences. While the semi-structured datasets consisted only of observations containing sampling metadata, the full dataset consisted of data gathered from diverse data sources; thus, unstructured data were the most common data type. The 2 goals of our study were (1) to examine whether it is possible to successfully model the distributions of species using only very small numbers of high-quality observations collected by volunteers and (2) to gain insights into the seasonal movements and habitat use of the species, validated by our expert knowledge on the species ecology.

## MATERIALS AND METHODS

### Study Species

The Plain Tyrannulet is a poorly known, medium-distance migratory flycatcher that inhabits the lowlands of southern South America (Ridgely and Tudor 1994; see Supplementary Material Appendix S1). It breeds in the hot and dry Chaco region of Argentina, Paraguay, and Bolivia where it inhabits woodlands and forest edges (Fitzpatrick 2020). However, breeding data come from just 5 observations (see Zyskowski et al. 2003, Di Giacomo 2005, Bodrati 2019, Fitzpatrick 2020, J. I. Areta personal observation) leading to a poor understanding of its breeding range. During the austral winter, it migrates north to the Pantanal and southwestern Amazonia where it overwinters mostly in riparian habitats and early successional vegetation (Chesser 1995, Stotz et al. 1996). The tyrannulet can be found in reduced numbers during the winter in northern Argentina (Coconier et al. 2007, Capllonch et al. 2009, Pearman and Areta 2020, 2021), suggesting that the species is a partial migrant with partially overlapping breeding and nonbreeding ranges.

## Data

**Data sources.** We gathered observations of Plain Tyrannulet from community science projects and web platforms such as eBird (ebird.org), Macaulay Library (macaulaylibrary.org), Xeno-Canto (xeno-canto.org), EcoRegistros (ecoregistros.org), and iNaturalist (inaturalist.org) that were available as of December 2019. We obtained additional data from published ornithological works listing Plain Tyrannulet, from digitized holdings in natural history museums, by personal examination of specimens, and contributed our unpublished data. Records were separated into those corresponding to the peak breeding season (November–February) and the peak nonbreeding season (May–August); thus, we discarded data from possibly transitional months (March–April and September–October). We kept records from only 1999 to 2019 to properly match our observational data with the available habitat-description data (see Environmental Variables later in this section). We only used records that had precise information on the location (i.e. referring to a specific site or point rather than an entire area or region) and date, not including records from WikiAves Brazil (wikiaves.com.br) that only have locations at the municipality/county level. Because misidentifications can bias phenological assessments (Gorleri and Areta 2021) and Plain Tyrannulets can be challenging to identify (Pearman and Areta 2021), we stringently vetted the datasets to ensure that no misidentified documented or implausible undocumented records were included.

**Datasets.** We created separate semi-structured datasets for the breeding and nonbreeding seasons of the Plain Tyrannulet to be compared with 2 other datasets that contained the full set of detection records of the species for each of these 2 seasons (Figure 1 and Supplemental Material Figure S1). By modeling with these datasets, we were able to investigate the relative accuracy of distribution models based on a smaller dataset of semi-structured records in which variation in the observation process is measured, compared with models built with a larger number of unstructured records in which variation in the observation process is undescribed. The smaller dataset (semi-structured dataset) consisted of detection records and inferred non-detection locations of the tyrannulet obtained from eBird (i.e. complete checklists in which Plain Tyrannulet was not reported) for which sampling effort metadata were available. The other 2 datasets contained all available records of the tyrannulet, but in one case with the inferred non-detection locations obtained from eBird simulating a presence–absence dataset (unstructured PA dataset) and in a second case combined with randomly generated background points (unstructured PB dataset), a common practice among researchers to model with unstructured, presence-only data.

To create the semi-structured dataset, we first filtered to detection records for which there were ancillary data that described the observation process (distance traveled, duration, and number of observers); this ancillary information was available only from eBird data. We filtered these semi-structured records to retain only those that met the following criteria: observations were made using either the "stationary" or "traveling" protocols, the duration of the observation period was no more than 5 hr, the distance traveled during the observation period was no more than 5 km, and 10 or fewer observers were in the group of observers. These post hoc filters create a set of more standardized surveys from the larger dataset (see Johnston et al. 2021). The semi-structured detection records included both the observations of the tyrannulets and the ancillary metadata that we used in our distribution models to account for sources of variation in detection rate.

Second, we obtained non-detection locations of the tyrannulet from eBird to analyze data in a presence–absence framework. Each eBird complete checklist—a list of species for which the observers have indicated that all detected species are recorded—was treated as a non-detection if Plain Tyrannulet was not recorded. The 2 advantages of our approach are that these non-detection records are at locations we know that observers have visited and so nonrandom site selection by observers is taken into account and that the non-detection records also contain information about the observation process that is identical to the ancillary information (i.e. distance traveled, duration, and number of observers) in the semi-structured detection records. To cover the entire environmental gradient of the distribution of the species, we used non-detection records within the potential distribution range of Plain Tyrannulet for each of the breeding and nonbreeding seasons. To do this, we created a circular buffer of 100 km around each detection point, and then we generated a minimum convex polygon over the resulting buffer that we used as a bounding box for the extraction of non-detections. We decided to use this radius instead of a narrower one (e.g., 50 km) given that the full geographic range of the Plain Tyrannulet is yet to be fully understood, and it is not unlikely that the species occurs 100 km away from many of the current edge-records that exist. We further filtered to non-detection records that met the same sampling criteria as semi-structured detection records. Finally, we joined both the semi-structured detection and non-detection datasets. The resulting semi-structured dataset consisted of 94 detections, and 9,803 non-detections for the breeding season (Figure 1), and 208 detections, and 21,708 non-detections for the nonbreeding season (Supplementary Material Figure S1), all of them associated with sampling effort metadata.

To create the unstructured PA datasets, we joined the full set of detection records of the tyrannulet obtained from

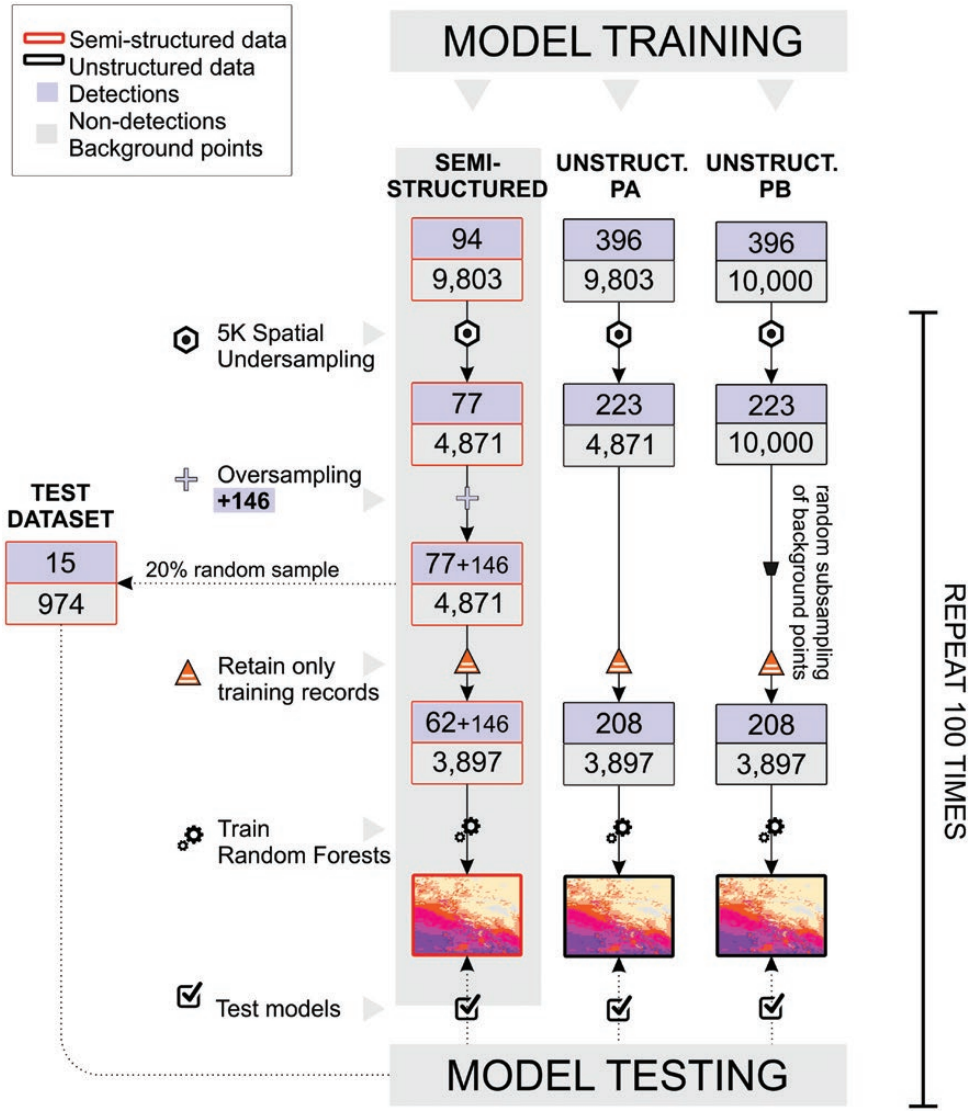


**FIGURE 1.** Flow diagram for the processing of Plain Tyrannulet (*Inezia inornata*) data into breeding season models' training and test sets. The semi-structured dataset consisted of detections and inferred non-detections, both with sampling metadata. Unstructured datasets consisted of the full set of detections associated with inferred non-detections (PA suffix) or with background points (PB suffix), none with sampling metadata. Inside each box, we indicate the number of detections (lilac-filled boxes) and non-detections or background points (gray-filled boxes). The red frame indicates that data are semi-structured (i.e. including data sampling metadata), and the black frame indicates unstructured data. See flow diagram for nonbreeding models in Supplementary Material Figure S1.

diverse sources, including eBird records, to the inferred non-detection locations obtained from eBird. Because of the diverse nature of detection records used in this dataset, metadata describing the observation process were available only from semi-structured records. However, we removed these metadata to match in format the more abundant unstructured data. We decided to create this presence–absence dataset because recent studies have demonstrated that models using presence-only data may benefit if actual non-detections are used instead of randomly generated background points as surrogates of the species absences (Henckel et al. 2020, Johnston et al. 2021). The resulting unstructured PA dataset consisted of 396 detections, and 9,803 non-detections for the breeding season (Figure 1), and 786 detections, and 21,708 non-detections for the nonbreeding season (Supplementary Material Figure S1), all of them without sampling effort metadata.

Finally, to create the unstructured PB datasets, we joined the full set of detection records of the tyrannulet to 10,000 randomly generated background points. This procedure creates a presence-background dataset, which is a commonly used data format for species distribution modeling when data are stored as presence-only (Barbet-Massin et al. 2012). We generated the background points inside the

same minimum convex polygons that we used to extract the non-detection locations for each season (see above in this section). We performed this procedure using the Data Management tool in ArcMap v10.4 (ESRI, Redlands, California). The resulting unstructured PB dataset consisted of 396 detections and 10,000 background points for the breeding season (Figure 1), and 786 detections and 10,000 background points for the nonbreeding season (Supplementary Material Figure S1), none with sampling effort metadata.

**Environmental variables.** We modeled the distributions of Plain Tyrannulets as a function of habitat, topography, and climate. The habitat data were from MODIS land cover product MCD12Q1 v006 (Friedl and Sulla-Menashe 2015), and we computed the percentage of each land cover class that was present within a 2.5 × 2.5 km square region (i.e. 5 MODIS pixels) centered on each location for the year in which the sighting was submitted. Because the habitat layers of this landcover product date from the year 2001 to 2018, we used 2001 landcover data for records dating from the years 1999 and 2000, and likewise, we used 2018 landcover data for records dating from the year 2019. We removed the land cover of class 1 (Evergreen Needleleaf Forests) and class 3 (Deciduous Needleleaf Forests) because these habitat types are absent across the study area. Our climate data were monthly rainfall and monthly temperature from WorldClim v2.1 (Fick and Hijmans 2017). We related each record to the average rainfall and average temperature values for the month in which the record was submitted. Additionally, we added 2 time-invariant covariates: elevation, from the digital elevation model CGIAR-CSI SRTM 90m 4.1 (30 arcsec, ~1 km) (Jarvis et al. 2008), and Euclidean distance to rivers, using the Rivers and Lake Centerlines layer from Natural Earth products 4.1.0 (Natural Earth 2020). We standardized to 2.5 min, the resolution of variables other than habitat using the R package *raster* (Hijmans and van Etten 2016). The full list of predictor variables is provided in Supplementary Material Table S1.

**Spatial undersampling.** Opportunistically collected data tend to be non-randomly distributed in space and time, and therefore spatial bias needs to be accounted for before analysis (Guillera-Arroita et al. 2015). To ameliorate the effect of spatial bias, we conducted spatial filtering of the detection and non-detection records (see Robinson et al. 2018). We created a hexagonal grid across the breeding and nonbreeding areas with 5 km between the centers of adjacent hexagons, using the R package *dggridR* (Barnes and Sahr 2017). We chose to use hexagons rather than squares because hexagons provide less spatial distortion (Sahr 2011). Then, we randomly selected one detection and one non-detection from each hexagon from each week within years, if a record of that type was present. By doing this, we ensured the spatiotemporal independence of each record to reduce the chance that we selected overlapping records of the same class for each model.

In addition, by separately selecting detections and non-detections, we achieved a closer balance between detection and non-detection records (Figure 1), a procedure that has been demonstrated to improve inferences with highly imbalanced data (Robinson et al. 2018). As the spatial undersampling randomly chooses one detection and one non-detection within a grid cell, many records may be excluded from the training dataset; therefore, we repeated this undersampling process before splitting data into test and training datasets for each of multiple iterations of model fitting for each dataset. Note that we did not perform spatial undersampling on the background points since they were already randomly distributed in the geographic space.

**Oversampling.** We ensured that any observed improvement in model accuracy was not simply a function of increasing the sample size or variation in the proportion of detections across the datasets. Model accuracy can increase with larger sample sizes (e.g., Stockwell and Peterson 2002), and the proportions of detection and non-detection records are important because models will preferentially predict the more common group with higher accuracy (McPherson et al. 2004). Differences in sample sizes and the proportions of the 2 classes can affect some of the metrics that we used to assess model accuracy (Longadge et al. 2013). Our semi-structured dataset had fewer detections than the unstructured datasets because it contained only a subset of all available records (Figure 1). Therefore, to remove the effect of variation in the number of detections and non-detections across the datasets that we compared, we equalized sample sizes to that of the unstructured dataset by oversampling detections from the semi-structured dataset. We used the synthetic minority oversampling technique (SMOTE, Chawla et al. 2002) to perform the oversampling procedure. Instead of creating exact copies of the existent data, SMOTE creates examples of the data that occupy the parameter space between a randomly chosen record and its nearest neighbor. This process resulted in datasets that had the same number of records and an equal proportion of detections and non-detections (Figure 1 and Supplemental Material Figure S1).

## Data Analysis

**Geographic coverage of detections.** Because the geographic coverage of data may affect the accuracy of a distribution model (Brotons et al. 2004), we evaluated whether detection records from the semi-structured dataset and the detection records from the unstructured datasets were similarly distributed in the geographic space for both breeding and nonbreeding seasons. We assessed this

by (1) counting the number of observations for each data type and (2) visualizing the core area of coverage, measured as the minimum convex polygon encompassing 95% of all the observations for each data type. This polygon was generated by removing 5% of records that were farther (Euclidean distance) from the median center of all records. We then calculated the area of each polygon to compare the overall geographic coverage of each data type.

**Species distribution models.** We selected and spatially subsampled 20% of the semi-structured detections (excluding oversampled observations) and 20% of the inferred non-detections as testing data for the evaluation of each model using cross-validation. This test set was selected because, first, all detection records are also contained in both unstructured datasets and, second, because actual non-detections are more appropriate than background pseudoabsences for testing models since they reflect true sites where the species was not detected. For the training datasets, we removed any observation that was in the test set. Also, we randomly selected background points to equal the number of non-detections used for training. We repeated this process of separating data into training and testing subsets 100 times, creating 600 unique datasets, against which our models were tested (300 for each season).

We used the R package *ranger* (Wright and Ziegler 2017) to build a Balanced Random Forest model for the breeding and nonbreeding seasons separately using the semi-structured datasets and both unstructured datasets. Balanced Random Forest is a modification of the Random Forest algorithm designed for imbalanced data (Chen et al. 2004). In this approach, each decision tree that makes up the random forest contains an equal number of randomly selected records of the majority class (here non-detections or background points) and the minority class (here detections). We grew 1,000 classification trees for each model type and set to 4 the number of variables, from which the model could select at each split for each tree (James et al. 2013).

Because the original prevalence of the species was modified when datasets were (1) subsampled and (2) oversampled, we needed to calibrate the results to avoid inflating the prevalence in the observed probability of detection. To create the necessary calibration curve, we predicted the detection probability on the 80% training set using the Balanced Random Forest model. We then built a binomial Generalized Additive Model (GAM) with the R package *scam* (Pya and Wood 2015). We used the real observations as the response variable and the detection probabilities as the predictor variable. We specified 4 degrees of freedom for each GAM and constrained the shape to be monotonically increasing.

**Model evaluation.** We evaluated the accuracy of the models using multiple predictive performance metrics, each of which describes a different aspect of a model fit. We used the test dataset to evaluate the mean squared error (MSE) between model predictions and true presence or absence in the test set. We also evaluated the ability of each model to rank positive observations higher than negative ones using the area under the curve (AUC; Fielding and Bell 1997). We evaluated each model's ability to simultaneously predict the presence or absence using Cohen's kappa (Cohen 1960). We also evaluated the sensitivity (true positive rate) and specificity (true negative rate). To visually assess the accuracy of the models, we produced distribution maps for each season, and we also calculated a ranking of variable importance and partial-dependence values that describe the relationship of each predictor variable with the response. We used the default functions with *ranger* to produce both variable importance and partial-dependence values (for further detail, see the R code). We assessed the consistency of maps and species–habitat relationships based on current knowledge of the species distribution and ecology and to our own field experience with Plain Tyrannulets.

## RESULTS

We compiled a total of 1,182 records of Plain Tyrannulet for the breeding and nonbreeding seasons. Only ~30% of all records had a semi-structured format available for models' training (i.e. providing information describing the observation process; Figure 1 and Supplementary Material Figure S1). While semi-structured data provided 62 and 124 detection records to train the semi-structured breeding and nonbreeding season models, the unstructured datasets contained 208 and 432 detection records, respectively (Figure 1 and Supplementary Material Figure S1). Unstructured data were not only more abundant, but they also had a wider geographic extent, as assessed by 95% minimum convex polygons, during both the breeding season (semi-structured: 680,051 km$^2$; unstructured: 1,179,191 km$^2$) and the nonbreeding season (semi-structured: 2,006,290 km$^2$; unstructured 2,414,070 km$^2$; Figure 2).

The map predictions of our models showed an obvious shift between breeding and nonbreeding seasons, indicating northward, but incomplete, post-breeding migration in the Plain Tyrannulet (Figure 3 and Supplementary Material Figure S2). Semi-structured and unstructured PA models estimated a high probability of encountering the species across the Chaco region during the breeding season, which essentially defined the main boundary of the species' breeding range. This pattern was not clearly recovered with unstructured PB models, which showed a patchy breeding distribution and overpredicted in areas where the species is either absent (e.g., Chile) or

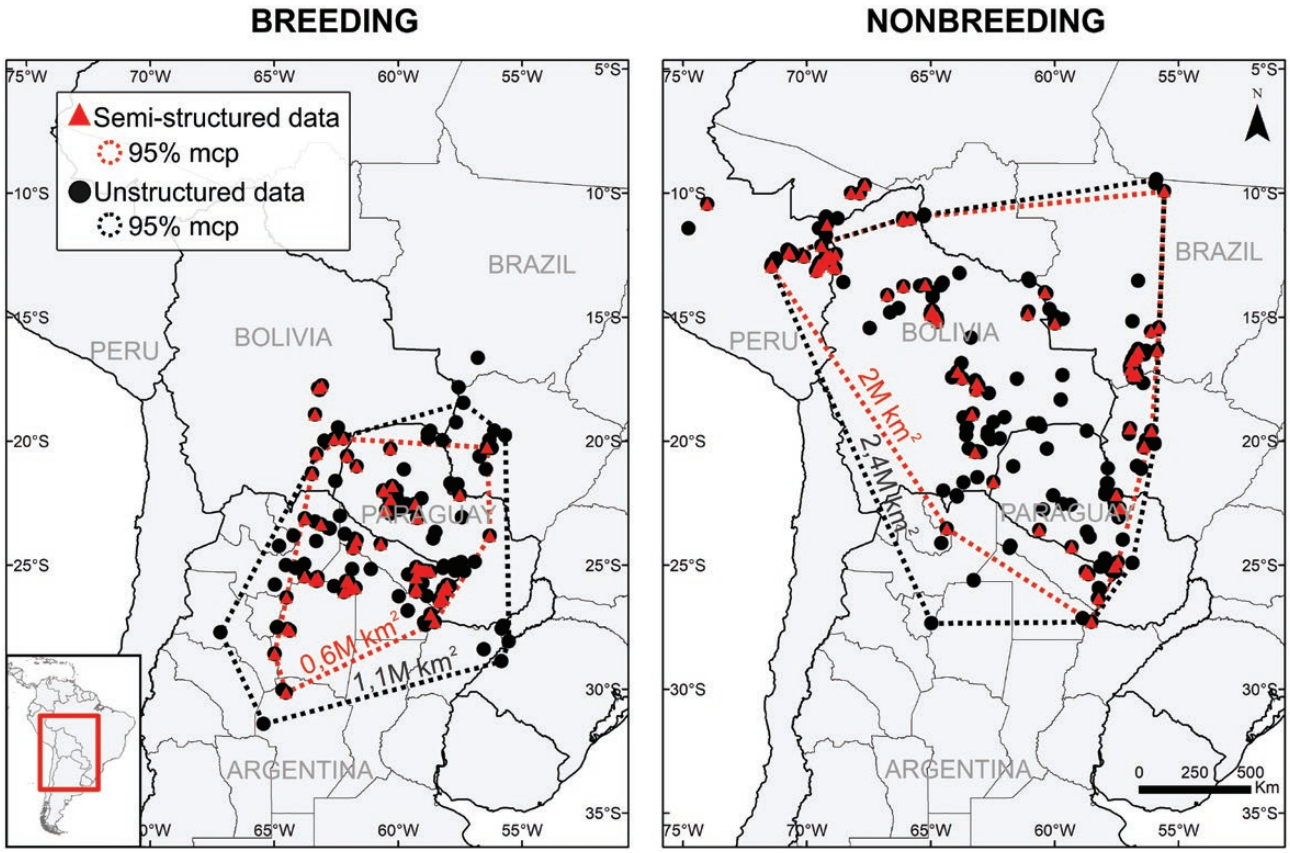


**FIGURE 2.** Geographic coverage of Plain Tyrannulet (*Inezia inornata*) detection data during the breeding and nonbreeding seasons. Red triangles indicate data that were collected using a semi-structured protocol, and black dots indicate those collected using an unstructured protocol. Dashed lines denote the minimum convex polygons (mcp) encompassing 95% of the observations of each class. Areas of polygons are indicated in millions (M) of square kilometers.

likely absent (e.g., eastern Uruguay and the adjacent region of Brazil). By contrast, all models generally coincided in the map predictions for the nonbreeding season. The regions with the highest encounter probability during the nonbreeding season were the savannas and wetlands in the Pantanal and the Beni and along rivers and semi-open areas in southwestern Amazonia (mostly in the Brazilian states of Acre and Rondônia).

Although semi-structured and unstructured PA models predicted similar distributions of Plain Tyrannulets, our quantitative assessment indicates that models built using semi-structured data showed a better predictive performance overall. This was a consistent result based on performance metrics from the 100 replicates of each model type, although some single iterations of data selection produced models that performed similarly well based on the measures of predictive performance (Figure 4). In the breeding season, semi-structured models outperformed unstructured models in all performance metrics (Figure 4). In the nonbreeding season, semi-structured models showed better accuracy in MSE, AUC, kappa, and specificity, with similar performance in terms of sensitivity relative to unstructured PA models (Figure 4). In addition, we found that unstructured PA models generally outperformed unstructured PB models, except for specificity in the nonbreeding season replicates (Figure 4).

We also found models based on semi-structured data to be more accurate in qualitative terms than the models based on unstructured data (see Discussion). All models broadly coincided in ranking the importance of predictors, with high predictive importance for temperature, elevation, and savannas for both seasons (Figure 5). The encounter probabilities were highest at a lower elevation, at higher temperature, and in savannas (Figures 6 and 7). This supports the notion that the species inhabits mostly semi-open habitats in hot lowlands. However, because the tyrannulet is an arboreal bird, semi-structured models were more accurate than unstructured-data models when associating tyrannulets more closely with habitats containing trees, such as woody savannas and deciduous forest in the breeding season, and evergreen forest in the nonbreeding season. In contrast, deciduous forests were not important in unstructured PA models even though the bird is found largely in this habitat type during the breeding season (Figure 5).

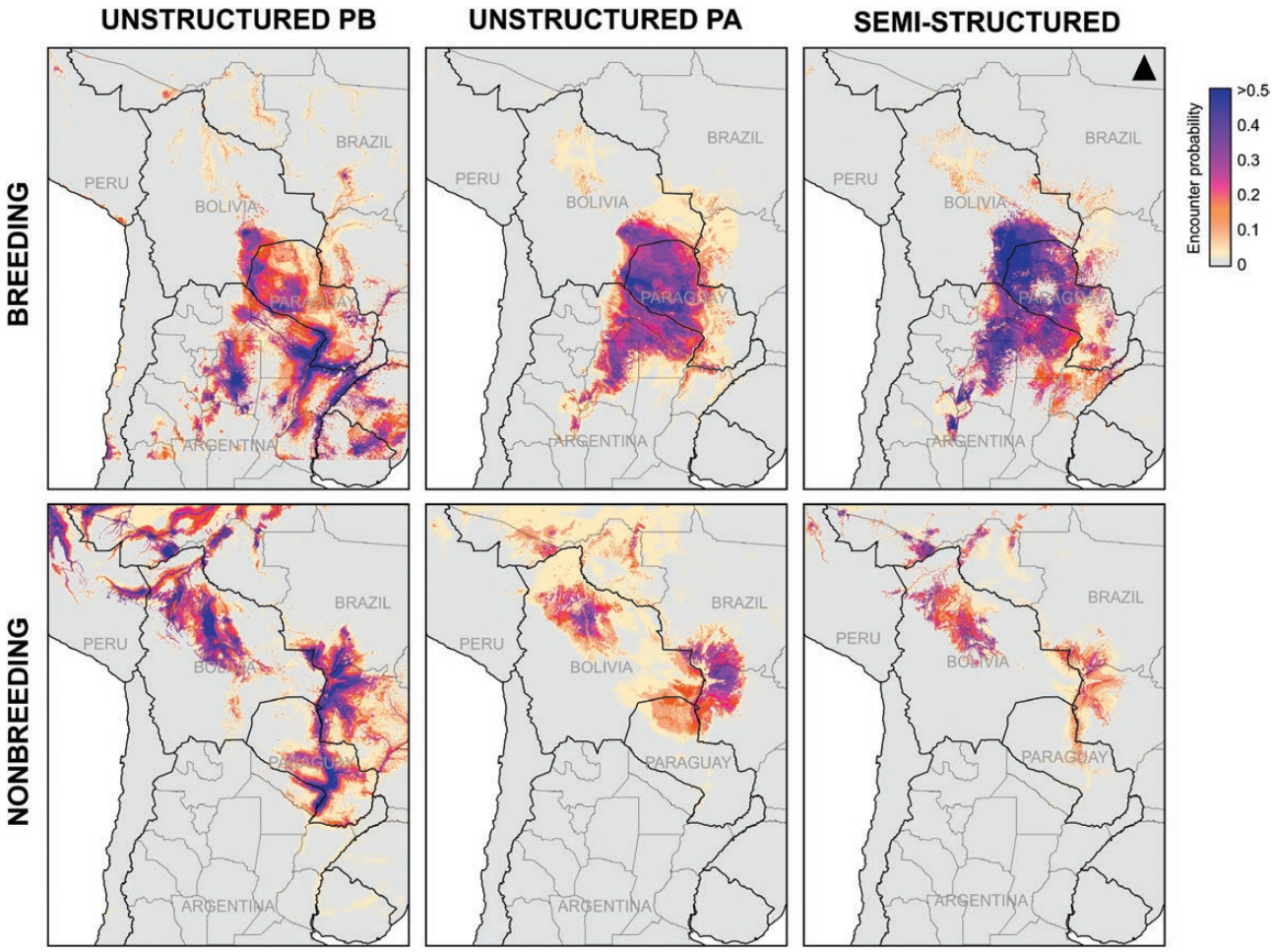


**FIGURE 3.** Estimated encounter probability of Plain Tyrannulet (*Inezia inornata*) for breeding (January 1) and nonbreeding (July 1) across the full geographic extent of the species distribution. The models from semi-structured data used detections and inferred non-detections, both with sampling metadata. Unstructured-data models used the full set of detections associated with inferred non-detections (PA suffix) or with background points (PB suffix), none with sampling metadata. For models using semi-structured data, encounter probabilities were calculated assuming a one-observer sampling lasting 2 hr, starting at 0600 hours and traveling a distance of 1 km. Maps are at a 2.5-min resolution. See full-resolution image in Supplementary Material Figure S2.

The partial-dependence plots obtained from semi-structured models were consistent with our own field experience and current evidence describing the environmental and habitat preferences of Plain Tyannulets. The exhibited response curves provided clear support that the tyrannulet is found in relatively dry areas during the breeding season, and near rivers in areas of higher precipitation during the nonbreeding season, avoiding areas with continuous *terra firme* forest. This was reflected by a shifting response to distance to rivers and to precipitation from breeding to nonbreeding season and a drop in the predicted probability of encountering Plain Tyrannulet at high proportions of evergreen forest in the landscape in the nonbreeding season (Figures 6 and 7). These patterns were not clearly recovered in the partial-dependence plots produced from the models using unstructured data. Moreover, unstructured-data models exhibited wrong or distorted habitat associations. For example, unstructured PB models erroneously related the tyrannulet to urban or grassland habitats (Figure 6), and unstructured PA models identified a spurious association with evergreen forest in the nonbreeding season model (Figure 7).

## DISCUSSION

We evaluated the utility of species distribution models for describing the seasonal range and environmental associations of the Plain Tyrannulet, a species for which the amount of available data and the information content provided by each record is presumably far from ideal for complex modeling. We demonstrated that the accuracy metrics and predictions of models improved when models were built with semi-structured data rather than with unstructured data, even when this entailed using a very

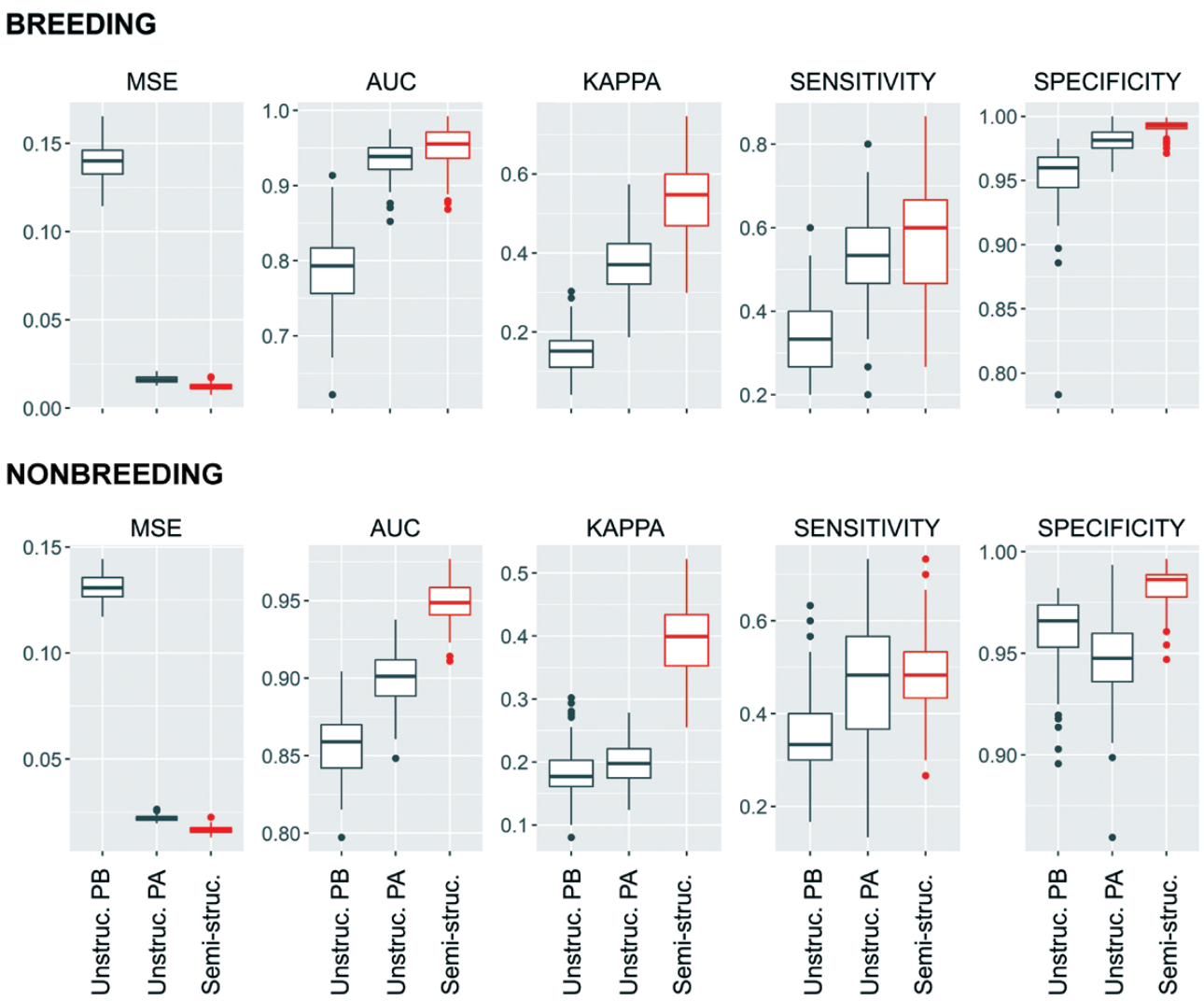


**FIGURE 4.** Accuracy metrics of the models created with the different datasets describing Plain Tyrannulet (*Inezia inornata*) breeding and nonbreeding distributions. The semi-structured dataset consisted of detections and inferred non-detections, both with sampling metadata. Unstructured datasets consisted of the full set of detections associated with inferred non-detections (PA suffix) or with background points (PB suffix), none with sampling metadata. Because each model type was run 100 times, each using a different randomly drawn dataset, summary distributions of the metric values are presented as boxplots.

small dataset of observations: only 62 of 208 detections for training the breeding season models and 124 of 432 for training the nonbreeding season models. The use of semi-structured data, therefore, promises to be of wide applicability to produce more reliable insights into distributional and ecological patterns even for data-poor bird species.

Including information on the observation process when modeling the distribution of a species leads to more accurate and robust results (Bird et al. 2014, Isaac et al. 2014, Milanesi et al. 2020, Johnston et al. 2021). We attribute the higher performance of semi-structured models relative to unstructured models of the Plain Tyrannulet to the existence of sampling metadata that we used as covariates for modeling. The improvement in models built with semi-structured data was clearly visible in (1) the higher predictive accuracy metrics and (2) the more accurate species–habitat relationships identified by the models. This later coinciding with current knowledge suggests that Plain Tyrannulets prefer semi-open and dry habitats during the breeding season (Bodrati 2004, 2019, Di Giacomo 2005, Areta and Gorleri personal observation) and wet savannas and riverine vegetation during the nonbreeding season (Chesser 1995, Stotz et al. 1996, Areta and Gorleri personal observation). While the overall geographic patterns described by all models were similar, semi-structured models showed a better contrast between areas of potential presence or absence of the species, resembling threshold effects. This can be explained by the sharper environmental

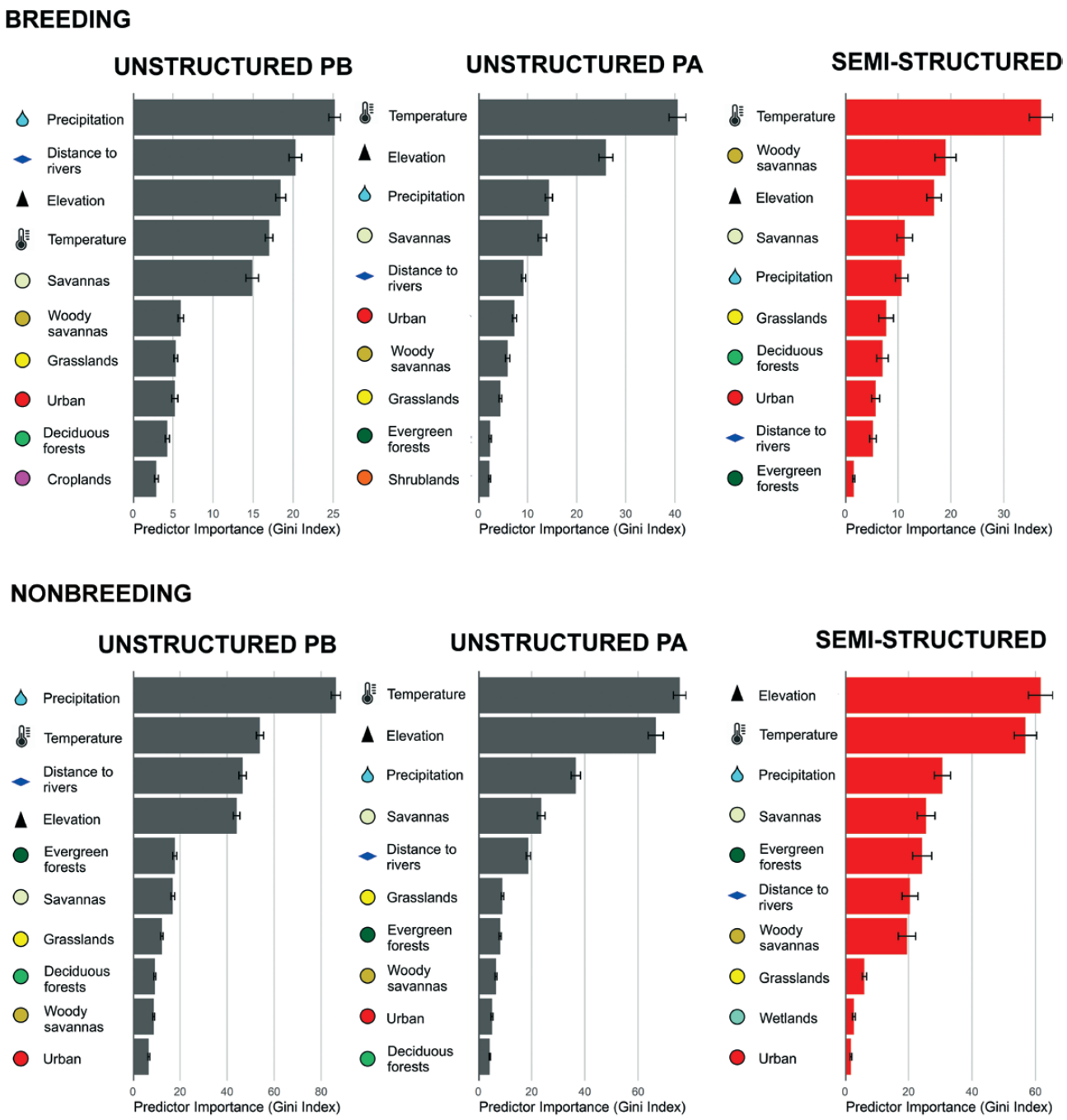


**FIGURE 5.** Important environmental and habitat predictors identified by models created with the different datasets of Plain Tyrannulet (*Inezia inornata*) for the breeding and nonbreeding seasons. The semi-structured dataset consisted of detections and inferred non-detections, both with sampling metadata. Unstructured datasets consisted of the full set of detections associated with inferred non-detections (PA suffix) or with background points (PB suffix), none with sampling metadata.

gradients identified by semi-structured models, which appear to be a general feature of models in which variation of detection rates is taken into account (Johnston et al. 2021).

Our findings are consistent with Steen et al. (2019) and Johnston et al. (2021), who also found improved accuracy of distribution models built with semi-structured data by only retaining high-quality records capable of accounting for data biases, even at the expense of reducing the number of observations for data analysis. While they assessed the effectiveness of this process with relatively common North American species with ample data, we demonstrated that this approach is also effective with a data-poor species, for

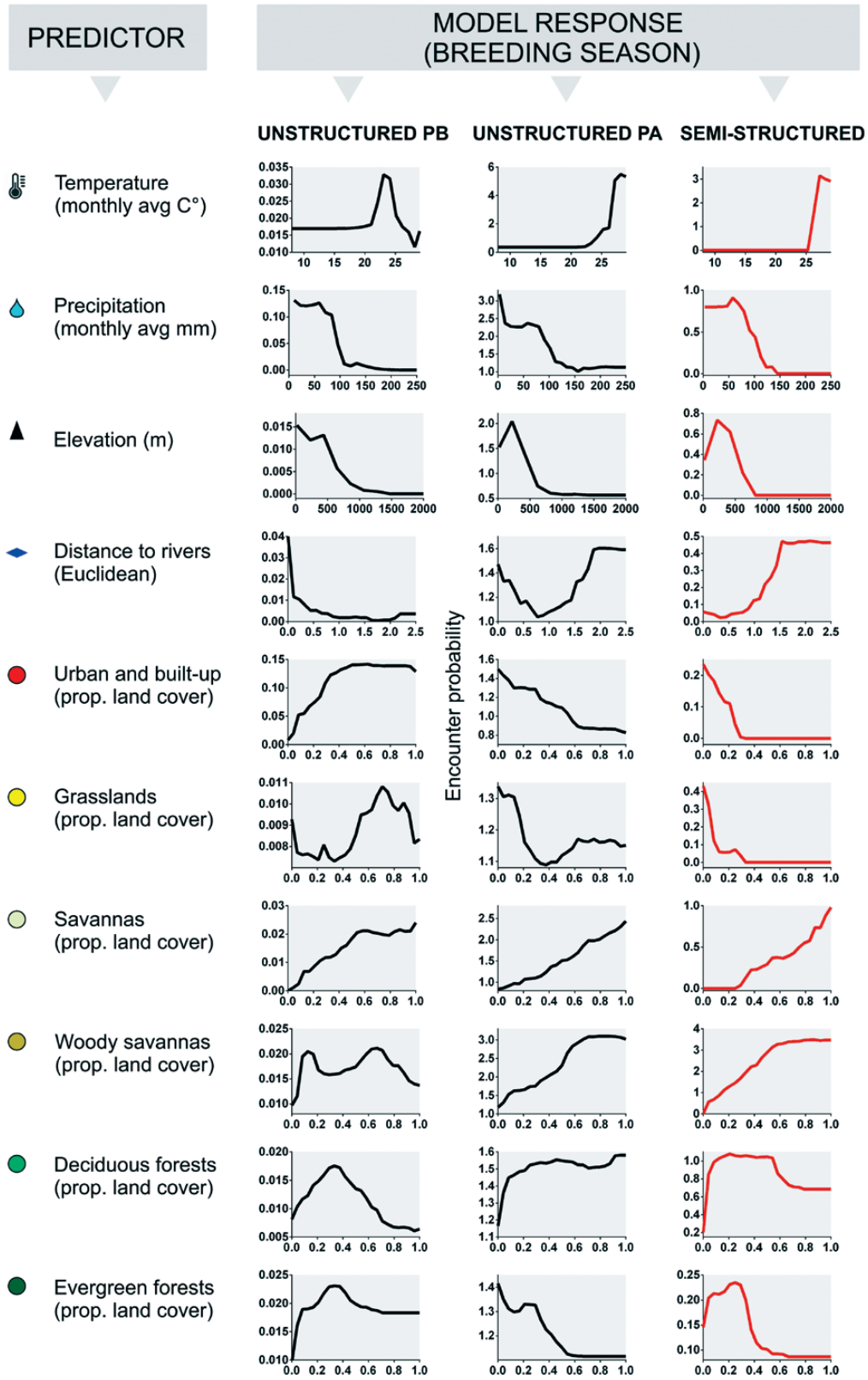


**FIGURE 6.** The partial dependence of encounter probability on the most relevant environmental and habitat predictors (Figure 5) identified by models created with the different datasets of Plain Tyrannulet (*Inezia inornata*) for the breeding season. The semi-structured dataset consisted of detections and inferred non-detections, both with sampling metadata. Unstructured datasets consisted of the full set of detections associated with inferred non-detections (PA suffix) or with background points (PB suffix), none with sampling metadata.

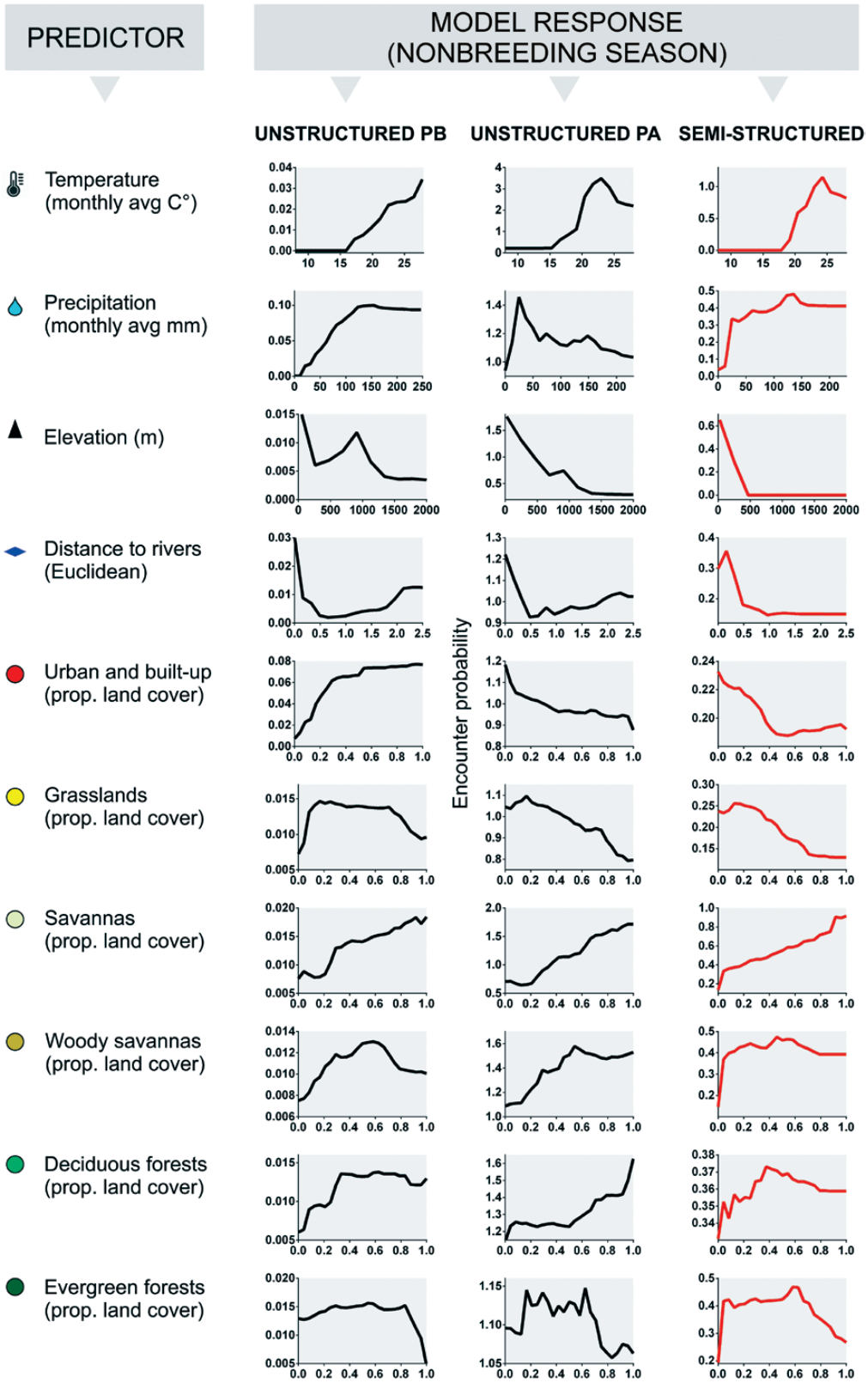


**FIGURE 7.** The partial dependence of encounter probability on the most relevant environmental and habitat predictors (Figure 5) identified by models created with the different datasets of Plain Tyrannulet (*Inezia inornata*) for the nonbreeding season. The semi-structured dataset consisted of detections and inferred non-detections, both with sampling metadata. Unstructured datasets consisted of the full set of detections associated with inferred non-detections (PA suffix) or with background points (PB suffix), none with sampling metadata.

which data filtering resulted in a substantial reduction in the absolute number of samples and geographic coverage of records (Figures 1 and 2). In the Plain Tyrannulet case, degrading semi-structured to unstructured data resulted in poorer performance of models. Thus, we echo Johnston et al. (2021) in recommending that data should not be degraded merely to obtain a set of data with a greater number of records of a species.

We also found that unstructured models of Plain Tyrannulet improved when using inferred non-detection records instead of background pseudoabsences, mirroring the findings of Henckel et al. (2020). Background data are typically used in combination with presence-only data for modeling (e.g., with MaxEnt); however, the resulting outputs are limited given the lack of true absence data (Brotons et al. 2004, Fithian et al. 2015). Our work provides further support that it is preferable to approximate non-detections instead of using random background data if constrained to work with presence-only records (Yackulic et al. 2013, Guillera-Arroita et al. 2015). In cases when data are of presence-only format, alternative methods exist for inferring non-detection locations (see van Strien et al., 2013, Milanesi et al. 2020).

Although our semi-structured models performed well with a partially migratory, poorly known, and relatively difficult to identify species, the Plain Tyrannulet, we acknowledge that the relative performance of models based on different types of data may vary among data-poor taxa. Before relying on semi-structured data to model the distribution of a data-poor species, it is critical to examine the spatial extent of the available data. Whereas our semi-structured dataset of Plain Tyrannulet had limited geographic coverage, the core breeding and nonbreeding ranges were properly represented by the data, and thus, the predictions of our models were still accurate (Figures 2 and 3). However, this condition may not be achieved in other data-poor species having large spatial data gaps. We strongly recommend mapping all available unstructured data of the species of interest before modeling, as this often illuminates seasonal and geographical patterns in data-poor taxa (Areta and Juhant 2019). Such information obtained from the more abundant unstructured data may serve as an external model validation or even to decide whether the use of only a subset of the available higher-quality data is appropriate for modeling or not.

The availability of semi-structured data, even in relatively small quantities, can improve our knowledge about the distribution of Neotropical bird species. As with Plain Tyrannulets, distributional knowledge of many other Neotropical migrant birds is coarse-grained (Faaborg et al. 2010, Jahn et al. 2020), which limits our ability to assess their conservation status. Our models from semi-structured data not only provided the first precise breeding and nonbreeding distribution maps for Plain Tyrannulets (Figure 3 and Supplementary Material Figure S2 and Appendix S1); but, given their relatively high spatial resolution, they were also able to highlight potential areas of the decline of the tyrannulet. For example, in the breeding season a "hole" of low encounter probability was mapped in the central Paraguayan Chaco (Figure 3 and Supplementary Material Figure S2). While the species is presumably abundant in western Paraguay (Lesterhuis et al. 2018), it may be locally declining in sites where unmanaged deforestation is transforming suitable woodlands into unsuitable habitats such as pastures and crops (Pacheco et al. 2021). Several other species may be suffering similar threats, and high-quality community science data are an excellent resource for generating useful products to quickly identify these threats to address more targeted conservation strategies.

The cumulative effort of citizen scientists is redefining the way biodiversity is monitored (Callaghan et al. 2019), but there is still a strong bias in the application of data toward northern hemisphere taxa (Feldman et al. 2021). For example, eBird is constantly updating and refining species' full-annual ranges with the use of semi-structured data, and modeled products improve as more observations are submitted to the project (see eBird Status and Trends; Fink et al. 2020). However, these products are often designed for large datasets and limitations arise in regions with large spatial data gaps. If the ultimate goal is to serve biodiversity conservation at a global scale, then one of the biggest challenges in the field of ecology is to develop or adapt advanced modeling tools to be applicable in poorly sampled regions because many of these regions harbor the greatest diversities of organisms worldwide (e.g., the Neotropics and Paleotropics). With this work, we would like to encourage researchers to use the recently available high-quality community science data, together with traditional knowledge, to further improve predictions of bird distributions in regions with deficient sampling. The resulting information will be critical for an effective conservation assessment of the species that inhabit these regions.

## CONCLUSION

We have shown a study case in which models using a reduced number of carefully vetted semi-structured community-science records outperformed models using greater amounts of unstructured data for a little-known migratory Neotropical flycatcher. This suggests that our approach can be profitably used to model the distributions of other data-poor migratory or resident bird species. We stress, however, the importance of examining the spatial distribution of the available data from all sources before modeling, to assess whether modeling with semi-structured data alone is considered justified.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Ornithological Applications* online.

## ACKNOWLEDGMENTS

We would like to thank the following people for either providing feedback or support for the preparation of this manuscript: Máximo Carlos Gorleri, Luisa Margarita Murdoch, Fabricio Candia, Natalia Bareiro, Bruno Bareiro, Leandro Bareiro Guiñazú, José Luis Navarro, Edelweiss and Silvia Enggist, Diego Núñez, Yoshitharo Kuroki, Dionel Aguiar, Enzo Gonzalo Moreno, Alejandro G. Di Giacomo, Orin Robinson, Heliana Guirado, and members of the ECOSON Lab: Emiliano Depino, Matías Juhant, Ingrid Holzmann, Freddy Burgos, Juliana Benitez, Emilio Ariel Jordan, and Juan Amaya. Also, we would like to thank all citizen scientists, who generously share their observations with the scientific and ornithological communities, and all the collectors and museums that obtained and preserve specimens used in this study.

**Funding statement**: This study was funded by the National Scientific and Technical Research Council (CONICET).

**Ethics statement:** This research was conducted in compliance with the CONICET Values and Ethics Code.

**Author contributions**: F.C.G. and J.I.A. conceived the idea; F.C.G. and J.I.A. did the bibliographical search of records; F.C.G. curated the records; W.M.H. contributed resources and ideas for data analysis; F.C.G. performed the data analyses; F.C.G. and J.I.A. performed the biological evaluation of models; and F.C.G. wrote the manuscript which was substantially edited and improved by J.I.A. and W.M.H.

**Conflict of interest statement:** We have no conflict of interest to disclose.

**Data availability**: R scripts to perform the distribution models of this study are archived at http://doi.org/10.5281/zenodo.5091795 (Gorleri et al. 2021).

## LITERATURE CITED

Areta, J. I., and A. Bodrati (2010). Un sistema migratorio longitudinal dentro de la selva atlántica: Movimientos estacionales y taxonomía del Tangará Cabeza Celeste (*Euphonia cyanocephala*) en Misiones (Argentina) y Paraguay. Ornitología Neotropical 21:71–86.

Areta, J. I., and M. A. Juhant (2019). The Rufous-thighed Kite *Harpagus diodon* is not an endemic breeder of the Atlantic Forest: Lessons to assess Wallacean shortfalls. Ibis 161:337–345.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller (2012). Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution 3:327–338.

Barnes, R., and K. Sahr (2017). dggridR: Discrete Global Grids for R. R package version 2.0.4. doi:10.5281/zenodo.1322866

Biddle, R., I. Solis-Ponce, M. Jones, S. Marsden, M. Pilgrim, and C. Devenish (2021). The value of local community knowledge in species distribution modelling for a threatened Neotropical parrot. Biodiversity and Conservation 30:1803–1823.

Bird, T. J., A. E. Bates, J. S. Lefcheck, N. A. Hill, R. J. Thomson, G. J. Edgar, R. D. Stuart-Smith, S. Wotherspoon, M. Krkosek, J. F. Stuart-Smith, et al. (2014). Statistical solutions for error and bias in global citizen science datasets. Biological Conservation 173:144–154.

Bodrati, A. (2004). Aportes al conocimiento de la distribución, abundancia y hábitat del Piojito Picudo (*Inezia inornata*) en la región chaqueña. Nuestras Aves 48:10–11.

Bodrati, A. (2019). Apuntes sobre un nido del Piojito Picudo (*Inezia inornata*) en la región chaqueña de Argentina. Nuestas Aves 64:19–20.

Breiner, F. T., A. Guisan, A. Bergamini, and M. P. Nobis (2015). Overcoming limitations of modelling rare species by using ensembles of small models. Methods in Ecology and Evolution 6:1210–1218.

Brotons, L., W. Thuiller, M. B. Araújo, and A. H. Hirzel (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437–448.

Callaghan, C. T., J. J. L. Rowley, W. K. Cornwell, A. G. B. Poore, and R. E. Major (2019). Improving big citizen science data: Moving beyond haphazard sampling. PLoS Biology 17:1–11.

Capllonch, P., D. Ortiz, and K. Soria (2009). Migraciones de especies de Tyrannidae de la Argentina: Parte 2. Acta Zoológica Lilloana 53:77–97.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:321–357.

Chen, C., A. Liaw, and L. Breiman (2004). Using Random Forest to Learn Imbalanced Data. University of Berkeley Technical Report 666, University of California, Berkeley, CA, USA. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

Chesser, R. T. (1994). Migration in South America: An overview of the austral system. Bird Conservation International 4:91–107.

Chesser, R. T. (1995). Biogeographic, ecological, and evolutionary aspects of South American austral migration, with special reference to the family Tyrannidae. Historical Dissertations and Theses, Louisiana State University, Baton Rouge, LA, USA.

Coconier, E. G., I. Roesler, F. Moschione, M. Pearman, P. Blendinger, A. Bodrati, and D. Monteleone (2007). Lista Comentada De Las Aves Silvestres de la Unidad de Gestión Acambuco. In Las Aves Silvestres de Acambuco. Temas de Naturaleza y Conservación (E. G. Coconier, Editor). Aves Argentinas/Asociación Ornitológica del Plata, Buenos Aires, Argentina. Vol. 6:32–103.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37–46.

Da Silveira, N. S., M. H. Vancine, A. E. Jahn, M. A. Pizo, and T. Sobral-Souza (2021). Future climate change will impact the size and location of breeding and wintering areas of migratory thrushes in South America. Ornithological Applications 123:2.

DeGroote, L. W., E. Hingst-Zaher, L. Moreira Lima, J. Whitacre, J. B. Slyder, and J. W. Wenzel (2020). Citizen science data reveals the cryptic migration of the Common Potoo *Nyctibius griseus* in Brazil. Ibis 163:380–389.

Di Giacomo, A. G. (2005) Aves de la Reserva El Bagual. In Historia Natural y Paisaje de la Reserva El Bagual, provincia de Formosa, Argentina. Temas de Naturaleza y Conservación (A. G. Di Giacomo and S. F. Krapovickas, Editors). Aves

Argentinas/Asociación Ornitológica del Plata, Buenos Aires, Argentina. Vol. 4:201–465.
Elith, J., and J. R. Leathwick (2009). Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.
Faaborg, J., R. T. Holmes, A. D. Anders, K. L. Bildstein, K. M. Dugger, S. A. Gauthreaux, P. Heglund, K. A. Hobson, A. E. Jahn, D. H. Johnson, et al. (2010). Recent advances in understanding migration systems of New World land birds. Ecological Monographs 80:3–48.
Feldman, M. J., L. Imbeau, P. Marchand, M. J. Mazerolle, M. Darveau, and N. J. Fenton (2021). Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. PLoS One 16:e0234587.
Fick, S. E., and R. J. Hijmans (2017). WorldClim 2: New 1 km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37:4302–4315.
Fielding, A. H., and J. F. Bell (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38–49.
Fink, D., T. Auer, A. Johnston, M. Strimas-Mackey, O. Robinson, S. S. Ligocki, W. Hochachka, C. Wood, I. Davies, M. Iliff, et al. (2020). eBird Status and Trends. Data Version: 2019. Released: 2020. Cornell Lab of Ornithology, Ithaca, New York, USA. https://ebird.org/science/status-and-trends
Fithian, W., J. Elith, T. Hastie, and D. A. Keith (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6:424–438.
Fitzpatrick, J. W. (2020). Plain Tyrannulet (*Inezia inornata*), version 1.0. In Birds of the World (J. del Hoyo, A. Elliott, J. Sargatal, D. A. Christie, and E. de Juana, Editors). Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bow.platyr1.01
Friedl, M., and D. Sulla-Menashe (2015). MCD12Q1 MODIS/Terra + Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
Gorleri, F. C., and J. I. Areta (2021). Misidentifications in citizen science bias the phenological estimates of two hard-to-identify *Elaenia* flycatchers. Ibis. doi:10.1111/ibi.12985
Gorleri, F. C., W. Hochachka, and J. I. Areta (2021). Data from: Distribution models using semi-structured community science data outperform unstructured-data models for a data-poor species, the Plain Tyrannulet. Ornithological Applications 123:4. doi:10.5281/zenodo.5091795
Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. Mccarthy, R. Tingley, and B. A. Wintle (2015). Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24:276–292.
Guisan, A., and N. E. Zimmerman (2000). Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.
Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter (2013). Big data and the future of ecology. Frontiers in Ecology and the Environment 11:156–162.
Hayes, F. E., B. D. Hayes, and P. Lecourt (2018). Seasonal distribution of the Striated Heron (*Butorides striata*) in Southern South America: Evidence for partial migration. Hornero 33:105–111.
Henckel, L., U. Bradter, M. Jönsson, N. J. B. Isaac, and T. Snäll (2020). Assessing the usefulness of citizen science data for habitat suitability modelling: Opportunistic reporting versus sampling based on a systematic protocol. Diversity and Distributions 26:1276–1290.
Hijmans, R. J., and J. van Etten (2016). raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster
Isaac, N. J. B., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 5:1052–1060.
Jahn, A. E., V. R. Cueto, C. S. Fontana, A. C. Guaraldo, D. J. Levey, P. P. Marra, and T. B. Ryder (2020). Bird migration within the Neotropics. The Auk: Ornithological Advances 137:4.
James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). An Introduction to Statistical Learning, Vol. 6. Springer, New York, NY, USA.
Jarvis, A., H. I. Reuter, E. Nelson, and E. Guevara (2008). Holefilled SRMT for the globe Version 4. CGIAR-CSI SRTM 90m. http://srtm.csi.cgiar.org/
Johnston, A., W. Hochachka, M. Strimas-Mackey, V. Ruiz Gutierrez, O. Robinson, E. Miller, T. Auer, S. Kelling, and D. Fink (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. Diversity and Distributions 27:1265–1277.
Kelling, S., A. Johnston, A. Bonn, D. Fink, V. Ruiz-Gutierrez, R. Bonney, M. Fernandez, W. M. Hochachka, R. Julliard, R. Kraemer, et al. (2019). Using semi-structured surveys to improve citizen science data for monitoring biodiversity. BioScience 69:170–179.
La Sorte, F. A., C. A. Lepczyk, J. L. Burnett, A. H. Hurlbert, M. W. Tingley, and B. Zuckerberg (2018). Opportunities and challenges for big data ornithology. Condor: Ornithological Applications 120:414–426.
La Sorte, F. A., and M. Somveille (2020). Survey completeness of a global citizen-science database of bird occurrence. Ecography 43:34–43.
Lees, A. C. (2016). Evidence for longitudinal migration by a "sedentary" Brazilian flycatcher, the Ash-throated Casiornis. Journal of Field Ornithology 87:251–259.
Lees, A. C., and R. W. Martin (2014). Exposing hidden endemism in a Neotropical forest raptor using citizen science. Ibis 157:103–114.
Lesterhuis, A. J., D. B. Villafañe, H. E. Cabral Beconi, and V. B. Rojas Bonzi (2018). Guía de las Aves del Chaco Seco paraguayo. Guyra Paraguay, Asunción, Paraguay.
Longadge, R., S. S. Dongre, and L. Malik (2013). Class imbalance problem in data mining. Review. International Journal of Computer Science and Network 2:83–87.
McPherson, J. M., W. Jetz, and D. J. Rogers (2004). The effects of species' range sizes on the accuracy of distribution models: Ecological phenomenon or statistical artefact? Journal of Applied Ecology 41:811–823.
Milanesi, P., E. Mori, and M. Menchetti (2020). Observer-oriented approach improves species distribution models from citizen science data. Ecology and Evolution 10:12104–12114.
Natural Earth. (2020). Free vector and raster map data. http://naturalearthdata.com/
Pacheco, P., K. Mo, N. Dudley, A. Shapiro, N. Aguilar-Amuschastegui, P. Y. Ling, C. Anderson, and A. Marx (2021). Deforestation Fronts—Drivers and Responses in a Changing World. WWF, Gland, Switzerland.
Pearman, M., and J. I. Areta (2020). Birds of Argentina and the South-west Atlantic. Field Guide. Helm, London, UK.

Pearman, M., and J. I. Areta (2021). Field identification of some look-alike *Serpophaga* tyrannulets and Plain Inezia from Argentina. Neotropical Birding 28:28–33.

Prudic, K. L., K. P. McFarland, J. C. Oliver, R. A. Hutchinson, E. C. Long, J. T. Kerr, and M. Larrivée (2017). eButterfly: Leveraging massive online citizen science for butterfly conservation. Insects 8:53.

Pya, N., and S. N. Wood (2015). Shape constrained additive models. Statistics and Computing 25:543–559.

Reese, G. C., K. R. Wilson, J. A. Hoeting, and C. H. Flather (2005). Factors affecting species distribution predictions: A simulation modeling experiment. Ecological Applications 15:554–564.

Ridgely, R. S., and G. Tudor (1994). The Birds of South America, Vol. 2. University of Texas Press, Austin, TX, USA.

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

Sahr, K. (2011). Hexagonal discrete global GRID systems for geospatial computing. Archives of Photogrammetry, Cartography and Remote Sensing 22:363–376.

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

Stockwell, D. R. B., and A. T. Peterson (2002). Effects of sample size on accuracy of species distribution models. Ecological Modelling 148:1–13.

Stotz, D. F., J. W. Fitzpatrick, T. A. Parker III, and D. K. Moskovits (1996). Neotropical Birds: Ecology and Conservation. University of Chicago Press, Chicago, IL, USA.

Sullivan, B. L., C. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Sumner, S., P. Bevan, A. G. Hart, and N. J. B. Isaac (2019). Mapping species distributions in 2 weeks using citizen science. Insect Conservation and Diversity 12:382–388.

Sun, C., A. K. Fuller, and J. E. Hurst (2018). Citizen science data enhance spatio-temporal extent and resolution of animal population studies. bioRxiv, doi:10.1101/352708, preprint: not peer reviewed.

Van Eupen, C., D. Maes, M. Herremans, K. R. R. Swinnen, B. Somers, and S. Luca (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., C. A. Van Swaay, and T. Termaat (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. Journal of Applied Ecology 50:1450–1458.

Welvaert, M., and P. Caley (2016). Citizen surveillance for environmental monitoring: Combining the efforts of citizen science and crowdsourcing in a quantitative data framework. SpringerPlus 5:1890.

Wright, M. N., and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software 77:1–17.

Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant, and S. Veran (2013). Presence-only modelling using MAXENT: When can we trust the inferences? Methods in Ecology and Evolution 4:236–243.

Zyskowski, K., M. B. Robbins, A. T. Peterson, K. S. Bostwick, R. P. Clay, and L. A. Amarilla (2003). Avifauna of the Northern Paraguayan Chaco. Ornitologia Neotropical 14:247–262.