Small herbaria contribute unique biogeographic records to county, locality, and temporal scales

Travis D. Marsico[1,15]* (iD) , Erica R. Krimmel[2,3,]*, J. Richard Carter[4], Emily L. Gillespie[5,6], Phillip D. Lowe[4], Ross McCauley[7], Ashley B. Morris[8,9], Gil Nelson[10,11], Michelle Smith[10,12], Diana L. Soteropoulos[1,13], and Anna K. Monfils[14]

PREMISE: With digitization and data sharing initiatives underway over the last 15 years, an important need has been prioritizing specimens to digitize. Because duplicate specimens are shared among herbaria in exchange and gift programs, we investigated the extent to which unique biogeographic data are held in small herbaria vs. these data being redundant with those held by larger institutions. We evaluated the unique specimen contributions that small herbaria make to biogeographic understanding at county, locality, and temporal scales.

METHODS: We sampled herbarium specimens of 40 plant taxa from each of eight states of the United States of America in four broad status categories: extremely rare, very rare, common native, and introduced. We gathered geographic information from specimens held by large (≥100,000 specimens) and small (<100,000 specimens) herbaria. We built generalized linear mixed models to assess which features of the collections may best predict unique contributions of herbaria and used an Akaike information criterion-based information-theoretic approach for our model selection to choose the best model for each scale.

RESULTS: Small herbaria contributed unique specimens at all scales in proportion with their contribution of specimens to our data set. The best models for all scales were the full models that included the factors of species status and herbarium size when accounting for state as a random variable.

CONCLUSIONS: We demonstrated that small herbaria contribute unique information for research. It is clear that unique contributions cannot be predicted based on herbarium size alone. We must prioritize digitization and data sharing from herbaria of all sizes.

KEY WORDS   biodiversity collection; biogeography; herbarium; Index Herbariorum; natural history collection; North American Network of Small Herbaria; rare plant; Small Collections Network; specimen; voucher.

American Journal of Botany 2020 107(11): 1563; http://www.wileyonlinelibrary.com/journal/AJB © 2020 The Authors. American Journal of Botany published by Wiley Periodicals LLC on behalf of Botanical Society of America. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made. • **1**

Herbaria are critical components of biological research infrastructure. The *Index Herbariorum*, a comprehensive, worldwide, online inventory of herbaria and their holdings, reports 686 active herbaria in the United States of America (USA; Thiers, 2020). Collectively, these institutions serve as repositories for over 78 million specimens and represent the most extensive sampling of vascular and nonvascular plant biodiversity in the USA, as well as the only source of verifiable data on botanical biodiversity over time (Page et al., 2015; Heberling and Isaac, 2017; Thiers, 2020). Traditional research uses of herbarium specimens include type collections for species' names and references for taxonomy, systematics, floristics, and biogeography. Over time the uses have expanded to answer questions about invasive species, species range shifts, pollution trends, bioprospecting, etc. (Lavoie, 2013; Heberling and Isaac, 2017; Heberling at al., 2019; McCartha et al., 2019).

Herbarium specimens contribute to a diversity of research areas, and researchers utilize an expanding set of techniques and analyses that did not exist when the specimens were initially collected (Heberling et al., 2019). For example, it is only since 2001 that herbarium specimens have been used for molecular phylogenetic analysis (Ristaino et al., 2001; Lavoie, 2013). Biodiversity informatics is another field that brings new analytical methods to herbarium specimen data, e.g., species distribution modeling (SDM) to map biodiversity and predict response to climatic changes, in addition to complementing studies that assess extinction risk and determine conservation priorities (Guralnick and Hill, 2009; Bloom at al., 2018; Lughadha et al., 2018). Herbarium specimens are also being used to assess changes in phenology resulting from climate change (Miller-Rushing et al., 2006; Calinger et al., 2013; Davis et al., 2015; Park and Schwartz, 2015; Rawal et al., 2015; Pearse et al., 2017; Willis et al., 2017; Brenskelle at al., 2019; Pearson, 2019). Digital imaging has facilitated research at unprecedented scales via low-cost automated and semi-automated techniques for scoring morphological characteristics or analyzing color (Gehan and Kellogg, 2017; Soltis, 2017).

Herbaria are essential partners in myriad large-scale, data-driven research initiatives not only within the plant sciences, but also extending into ecology, human health, and economics (Gropp, 2003; Winker, 2004; Pyke and Ehrlich, 2010; Heberling and Isaac, 2017). Studies in disease ecology and public health cite publications that use herbarium data from aggregated biodiversity occurrence databases (Ball-Damerow et al., 2019), exemplifying the integral connection between biodiversity and human health. These diverse new uses enhance, rather than replace, the traditional role of herbaria as research infrastructure (Heberling and Isaac, 2017). In fact, over the last century, citation of herbarium specimens has substantially increased, underscoring the vital role that herbaria continue to play in the future of cross-disciplinary, integrative science (Heberling et al., 2019).

Many emergent research techniques benefit from having more specimen records accessible, and researchers are clamoring for data to fill spatial, taxonomic, and temporal gaps (Ariño et al., 2013; Lavoie, 2013). Ball-Damerow et al. (2019, p. 2) assert that "the biggest obstacle for biodiversity data users is obtaining records of sufficient quantity and quality for the region and taxonomic group of interest." In a literature review of works citing herbarium specimens published between 1933 and 2012, Lavoie (2013) found that the median number of specimens referenced for biogeographic or conservation-focused studies was >2800. Species distribution modeling is a specific example of an approach greatly improved by a larger sample of specimen records, which might come from a combination of continued collecting, more spatially distributed collecting, and better access to existing specimen data (Feeley and Silman, 2011; Ball-Damerow et al., 2019). The contribution of small herbaria to SDM was addressed by Glon et al. (2017) in a case study of the Fuireneae (Cyperaceae). Using a combination of digitized data from small and large collections, the authors showed that species-specific models inclusive of data from small herbaria resulted in more refined predictions of ecological niche and enhanced SDMs bridging geographic gaps.

Collection bias is another known challenge that can be addressed on spatial, temporal, trait, phylogenetic, and collector planes by including a large number of specimen records (Ward, 2012; Meyer et al., 2016; Daru et al., 2017; Soltis, 2017). Bias can be minimized by increasing not only the total number of specimens, but also the number of collections providing specimens (Soberon, 1999; Krishtalka and Humphrey, 2000). In a case study featuring a common insect taxon, Ferro and Flick (2015) found that they needed specimens from a minimum of 15 collections to build a reasonable distribution model.

Herbaria have a rich history both as regional collections and as large institutions with national or global foci. Of the 686 herbaria in the USA, only 13 hold in excess of 1 million specimens each, representing a collective 40 million specimens (Thiers, 2020). Thirty-five collections hold 450,000 specimens or more and represent a collective 54.7 million specimens (Thiers, 2020). This means that approximately 30% (23 million) of the nation's total herbarium specimens are held across the 651 collections with fewer than 450,000 specimens each, many of which have fewer than 100,000 specimens (Barkworth and Murrell, 2012; Thiers, 2020). The sheer number and vast geographic distribution of these herbaria contribute to their collective value as research infrastructure and provide resources to an active scientific community both within the USA and internationally (Barkworth and Murrell, 2012; Lavoie, 2013).

Small herbaria are often regional in scope and contain fewer specimens than larger herbaria with a national or global scope, and regional herbaria are frequently defined by an ecological or taxonomic specialty as well as a geographic focus (Monfils et al., 2020). These collections may receive less research access than larger herbaria, in part because of the logistical advantage of traveling to a handful of larger herbaria over many smaller herbaria, a pattern demonstrated by López and Sassone (2019) for herbaria in Argentina. Similar visitation patterns based on collection size have been reported for entomology collections (Cobb et al., 2019). In a survey of herbaria globally, Lavoie (2013) found that the 63 individual large herbaria with >1 million specimens each were accessed three to six times more frequently than those with fewer specimens. However, although individual small (<100,000 specimens; 407 collections) and medium (100,000–999,999 specimens; 263 collections) herbaria were consulted less frequently, they collectively received a roughly equal number of consultations per size class, with small herbaria at 31% of consultations, medium herbaria at 39%, and large herbaria at 30%. Lavoie (2013) interpreted this as evidence that, despite containing only a fraction of total specimens worldwide, small herbaria contain specimens of local or national importance. O'Connell et al. (2004) have a similar finding; in their assessment of herbarium specimens collected on National Park Service land, they found records from 78 institutions collected between 1890 and 1980, with specimen detection rates inversely

related to collection size and with relevant specimens most often held by collections geographically close to the region of interest.

The advent of specimen digitization means that the logistical advantage of large vs. small herbaria is diminished because a researcher can often identify specimens of interest without visiting or necessarily contacting each individual herbarium. Before 2004, the use of digitized collections was practically nonexistent in the herbarium literature, but in the intervening years, digital access has become common and has facilitated the use of many more specimens per study, from a median of 226 specimens in studies that did not access digital records to a median of 15,295 specimens in studies that did (Lavoie, 2013). As digitized specimen records become available online, they have an even broader reach. Ball-Damerow et al. (2019) noted that online species occurrence databases are most commonly used for studies on species distribution, species richness, taxonomy, conservation, and invasive species—all research themes that gained prominence long before digitization. These online species occurrence databases are democratizing access to herbarium specimens from collections that have previously been difficult to access due to location and/or staffing. In fact, Lavoie (2013) attributes the lag in publications using digitized specimen data, which have been available in part since the 1970s, to the lack of online accessibility.

For most of the last 200 years, access to specimen-based biodiversity records has depended primarily on researchers traveling to collections or curators shipping loans upon request. To save time and funds, researchers have often limited their investigations to large, well-known institutions and those with adequate resources to support loan management, potentially ignoring important specimens and data deposited in less accessible or discoverable institutions, which are often smaller (Casas-Marce et al., 2012). A baseline understanding of the relative scientific contributions of specimen data in variously sized herbaria is essential, especially in light of the recent advances in collections digitization and data mobilization catalyzed by the USA National Science Foundation's (NSF) Advancing Digitization of Biodiversity Collections (ADBC) program, and given the continuing loss of support for biodiversity collections of all types (Winker, 2004), including the potential loss of the specimens themselves.

It is widely recognized that our knowledge of biodiversity is far from complete, even on a coarse geographic scale (Sorrie and Weakley, 2001; Meyer et al., 2016). Several authors have expressed support for the importance of including small collections' data for understanding temporal and biogeographic diversity (Snow, 2005; Barkworth and Murrell, 2012; Lavoie, 2013; Glon at al., 2017). One recent publication, in particular, highlights the future importance of discoverability and digitization of regional collections (Lendemer et al., 2020). Here, we studied the extent to which the holdings of small herbaria, often regional in scope, contribute meaningfully to our knowledge of plant biogeography at geographic and temporal scales. This paper advances such understanding by quantifying the unique contributions made by collections of all sizes.

MATERIALS AND METHODS

Herbarium specimen data were sampled in eight of the 50 USA states (16%), based on locations of collaborating authors: Arkansas (AR), California (CA), Colorado (CO), Florida (FL), Georgia (GA), Michigan (MI), Tennessee (TN), and West Virginia (WV). Using a state-based approach is justified because floras that contain distribution and abundance data for species are often written or compiled in state-specific floras and by state agencies, such as natural heritage programs. The states included in this study span the nation and represent a range of sizes (geographically and by population) and endemism. Botanical history—including number of herbaria, number of total specimens, and collection effort within the state—also varies across these states.

For each state, plant species (or infraspecific taxa, "taxa" hereafter) were selected within each of four status categories: extremely rare (S1, typically representing ≤5 population occurrences), very rare (S2, typically representing 6–20 population occurrences), common native, and introduced. Due to differences in phytogeography and the historical emphasis on state-based plant projects, taxa were selected for this project within each state rather than across states, resulting in a compiled list of 320 taxa to sample (8 states × 4 status categories × 10 taxa; except WV, which had 8 taxa in the common native category and 12 taxa in the introduced category).

To identify sample taxa in the S1 and S2 status categories, we acquired state-level lists for tracking rare/threatened/endangered plants from state natural resource conservation agencies (see data sources in Appendix S1). We chose to select taxa separately for the S1 and S2 categories because we wanted to analyze the occurrence of rare species records but were concerned that S1 taxa may be too infrequently represented in the specimen data set. Taxa with dual listings (i.e., S1/S2 or S2/S3) were excluded from our selections. Ten S1 taxa within each state and 10 S2 taxa within each state were randomly selected from the state-level lists using a random number generator to identify a row in a spreadsheet (filtered by status, S1 or S2) that correlated to a taxon. Or, in cases where the state-level list was formatted for print, the random number generator identified a page number on which the first taxon matching the correct status (S1 or S2) was selected. Despite this slight variation in selection approach across state-level lists, each researcher ensured that taxa were selected randomly to avoid bias.

To identify sample taxa in the introduced status category, we acquired state-level lists for tracking introduced/invasive plant species; if a state did not maintain its own introduced/invasive species list, an analogous list from a neighboring state was used (see data sources in Appendix S1). Introduced taxa within each state were randomly selected from the state-level lists via the same methods as above. In states that included data about the level of invasive threat (CA, FL, GA, TN), the randomly selected introduced taxa were chosen from a subset of those species representing the highest threat level. In states lacking these data (AR, CO, MI, WV), the randomly selected introduced taxa were compiled without accounting for perceived threat level.

To identify sample taxa in the common native status category, we acquired state-level lists of all taxa known to occur within the state from checklists, atlases, floras, or databases (see data sources in Appendix S1). Common native taxa within each state were randomly selected from the state-level lists via the same methods as above. We discarded any selected taxon listed as rare or introduced, and a new random number was generated until the selected taxon was absent from these other lists.

This design resulted in a random sample of taxa across states and species statuses, which reduced overall bias. Based on the randomly selected sample of 40 taxa per state, we attempted to acquire specimen data from all herbaria located within each state during the summer and fall of 2014 (see Appendix S2 for a list of herbaria

contacted). We gathered specimen information from online databases when available and by contacting curators or collections managers when online data were not available. When data were not available digitally, we digitized de novo from specimen images or specimen loans and repatriated the transcribed data back to the collection. Coauthors were responsible for acquiring and collating data within their respective states.

Herbaria included in this study were categorized into two size classes of small (<100,000 specimens) and large (≥100,000 specimens). The 100,000-specimens cutoff classifies 85% of herbaria in the USA as small (Thiers and Ramirez, 2020) and is reflective of recent publications in the USA herbarium community (Lavoie, 2013; Glon et al., 2017). Emerging research suggests that a more appropriate cutoff would be <175,000 specimens (classifying 90% of herbaria in the USA as small; Thiers and Ramirez, 2020) based on the Jenks natural breaks classification method (Monfils et al., 2020). To be conservative in our estimates and conclusions, we maintained the more traditionally accepted 100,000-specimen cutoff for small vs. large herbaria in our primary analyses and discussion presented here, although to be comprehensive we have also provided alternative analyses for the 175,000-specimen cutoff.

For each specimen, at a minimum we recorded the catalog or accession number, taxon identification, state, county, locality (as transcribed from the specimen label), collector, and collection date. The data were collated and nominally cleaned to accomplish the research purposes of this project, e.g., date strings transformed into formatted dates, taxon names synonymized with current taxonomy, counties validated (see Appendix S3 for a data dictionary that briefly describes each field and any transformations applied). The collated data set consisted of 21,546 specimen records (see Data Availability statement with this article) and included records lacking our minimum data quality standards, which were flagged and later excluded during analyses. The original data had varying degrees of cleanliness, but we did not fix additional issues (e.g., incompletely parsed locality information) that were beyond the scope of this research.

Specimen localities were georeferenced for spatial analysis. We used geographic coordinate information when available either in the original locality description (~7% of specimens, $N = 1454$) or from the herbarium database (~11% of specimens, $N = 2460$). Specimen localities without coordinates were georeferenced automatically using the GeoLocate API with OpenRefine (~73% of specimens, $N = 15,807$; Rios, 2019; OpenRefine Core Team, 2018). We georeferenced specimens for which GeoLocate could not automatically determine coordinates using the online GeoLocate tool in combination with research on Google Maps (~5% of specimens, $N = 1068$). A small subset of specimens did not have enough information to georeference at a level of precision below county; we reviewed and flagged these as unable to be georeferenced (~4% of specimens, $N = 757$). All coordinate data were evaluated in QGIS (QGIS Development Team, 2019) to find instances in which the county recorded on the specimen label did not match the county identity based on coordinates. Mismatches occurred for ~2000 specimen records, and we refined these georeferences using the online GeoLocate tool in combination with research on Google Maps.

From the collated data set consisting of 21,546 specimen records, we reviewed and excluded 1366 records with specific data quality or scope issues, i.e., county information missing, multiple counties listed, specimens suspected to be cultivated, multiple herbaria listed (e.g., specimens of a small field station herbarium managed physically on site at a large herbarium), and/or herbarium located out of state. Among the 1366 specimens eliminated were records from two out-of-state herbaria (RM in Wyoming and SJNM in New Mexico), which had extensive holdings of Colorado material.

The data set was reviewed for duplicate specimens, and we assigned flags for categories of uniqueness using R (Bivand and Lewin-Koh, 2019; Bivand and Rundel, 2019; Bivand et al., 2019; R Core Team, 2019; Wickham et al., 2019; Zhu, 2019; see Data Availability statement for code). For our purposes, we conservatively defined duplicate specimens as those of the same taxon collected on the same date in the same county by the same collector. We suspected a priori that there may be a large number of duplicate specimens in our data set due to the tradition of field botanists and herbarium curators developing extensive and long-lasting specimen exchanges among institutions. Regardless of whether the duplicates were held within a single herbarium or across multiple herbaria, we only retained a single specimen from each set of duplicates shared within an herbarium size class and discarded all specimens belonging to duplicate sets that were shared between large and small herbaria. We categorized uniqueness into three primary scales at which a specimen may contribute novel spatiotemporal data to knowledge of a taxon: (1) a county record ("unique county"), (2) a record of a locality georeferenced as >1 km apart from any other locality ("unique locality"), or (3) a record of a distinct historical time from a previously sampled locality ("unique time", determined as a year/month/day previously unrepresented in the data). In our analyses, we only included the largest scale for which a specimen contributed uniquely. In other words, although any specimen flagged as a unique county by default also represents a unique locality and a unique time, we did not include unique county specimens in our analyses of unique locality or time contributions.

To determine whether herbarium size class (small, large) and/or species status (S1, S2, common native, introduced) were important in predicting specimen uniqueness, we created three sets of generalized linear mixed-effects models with a binomial logistic regression, one set for each of our three scales (county, locality, and temporal). We then conducted model selection on each set with an information-theoretic approach based on Akaike information criterion (AIC; Anderson and Burnham, 2002). For each scale, our candidate set consisted of a null model, individual fixed effects models, and a full model with each of the individual variables included as additive fixed effects. Uniqueness (1 for yes, 0 for no) at a given scale (county, locality, or temporal) was our response variable, and herbarium size and species status category were the fixed effects. State was treated as a random variable to account for our methods, which did not sample states comprehensively, but rather based on locations of the coauthors. We determined the best model in our candidate set by identifying which had the lowest ΔAIC value that was also less than 2. Modelling was conducted in the R programming language using the lme4 package (Bates et al., 2015; R Core Team, 2019; see Data Availability statement for code). We confirmed fit for each of our full models (unique county, unique locality, and unique time) and tested for collinearity by evaluating the variance inflation factors and Cramer's V values in R (Lenth, 2020; Navarro, 2015; Fox and Weisberg, 2019; see Data Availability statement for code).

RESULTS

One hundred thirty-eight herbaria contributed to our project, of which 26 had ≥100,000 specimens, representing large herbaria

(see Appendix S2). States ranged from having one large herbarium within the state (AR, GA) to having 10 large herbaria (CA). One hundred twelve herbaria represent small herbaria with <100,000 specimens, and states ranged from having 6 (WV) to 37 (CA) small herbaria. According to estimates of total herbarium size, specimens held by the large herbaria included in this study number 12,953,200 (87.5%), and specimens held by the small herbaria number 1,858,833 (12.5%). This proportion is similar to that of all United States herbaria recorded in *Index Herbariorum*, for which large herbaria hold a collective 68 million (87.2%) and small herbaria 10 million (12.8%) specimens (Thiers and Ramirez, 2020). Within the original data set of 21,546 specimen records collated for this project, large herbaria contributed 15,143 specimens (70.3%), and small herbaria contributed 6403 specimens (29.7%). After excluding rows with data quality issues and accounting for duplicate records (defined above in Materials and Methods), our data set was condensed to 16,348 records, each representing a unique collecting event. Most specimens (89% of those held by small herbaria and 83% of those held by large herbaria) were unduplicated, and duplicates were more likely to be distributed only between large herbaria than either only between small herbaria, or shared between large and small herbaria (Table 1).

Our primary analysis was conducted on a further reduced subset of these data ($N = 15,792$) by excluding an additional 137 records that were classified as a unique time by our flagging but that did not have a collecting day recorded. The relative contribution of specimens by herbarium size varied widely by state (Fig. 1), but small herbaria across all states contributed a larger percentage (30.7% of 15,792) of specimens to this study than expected based on their holdings (12.5% of total specimens are held by the small herbaria included in this study; Appendix S2). Patterns at each of our uniqueness scales (county, locality, temporal) also varied widely by state (Fig. 2; see Appendix S4 for the data used to generate this figure). Small herbaria in some states exhibited similarities between the proportion of records they contributed to the analysis data set and the proportion of records they contributed to certain uniqueness scales (compare Figs. 1 and 2). For example, small herbaria contributed nearly one half of the specimens for Arkansas (Fig. 1), and nearly half of the unique records at each uniqueness scale were provided by small herbaria (Fig. 2). As expected, there were greater unique contributions from the temporal scale than from locality or county and from the common native and introduced taxa than from the S1 and S2 taxa (Fig. 3A).

Modeling the effects of size class, species status, and state, and then comparing these models using AIC (Table 2) allowed us to parse high-level findings from the complexity of our results. We found that at all uniqueness scales (county, locality, temporal), the

full model was weighted 100%, meaning that it provided the best balance between fit and parsimony (Table 2). Our best model also did well fitting the observed data, which we used as a comparison to assess the validity of our models in predicting the probability that a specimen represents unique information at different biogeographic scales (compare Fig. 3B with 3C). Because of the way we analyzed our data, the probability of a specimen contributing uniquely at one of the scales is 1 (Fig. 3B, C). In other words, with duplicated specimens across herbarium size classes removed (2.6% of specimen records; Table 1), all specimens originating as unduplicated anywhere or duplicated within size class represent unique contributions at the county, locality, or temporal scale for a given herbarium size class. The probabilities of uniqueness predicted by our models (Fig. 3C) show that large herbaria are predicted to have nearly twice the probability of small herbaria to contribute unique county records, but only slightly greater probability than small herbaria to contribute unique locality records. Since the probabilities sum to 1 across the uniqueness scales, small herbaria are predicted to have a greater probability than large herbaria of providing unique records at the temporal scale (Fig. 3C).

To account for emerging research (see Monfils et al., 2020), we produced the same models for our data using a cutoff of 175,000 specimens to distinguish between small and large herbaria. We found that the full model at each uniqueness scale was again weighted 100% (results available in Appendix S5; also see Data Availability statement). These results indicate that the same factors are at play for explaining unique contributions of small herbaria, even if the cutoff for what constitutes a small herbarium is raised.

DISCUSSION

Our results show that herbaria house primarily unduplicated specimens within their states, and they represent unique knowledge at all biogeographic scales (county, locality, temporal). Our findings demonstrate that research requiring a complete picture of existing biogeographic knowledge at any scale must include specimens from both small and large herbaria. Although previously it has not been widely demonstrated that small herbaria curate unduplicated specimens, we found that 97.4% of small herbarium specimens sampled for this study are either totally unduplicated, or duplicated only by another small herbarium (Table 1). These unduplicated specimens represent unique biogeographic knowledge in all species categories (Fig. 3A, B), and our models predict how this uniqueness is distributed across biogeographic and temporal scales (Fig. 3C). We show that within a given size class (small, large) and species status (S1, S2, common native, introduced), a specimen has an increasing probability of representing uniqueness at the county vs. locality vs. temporal scale. For example, our models predict that an unduplicated specimen from a small herbarium of an S2 taxon has approximately a 10% chance of representing a unique county, a 26% chance of representing a unique locality (additive with unique county contribution), and a 100% chance of representing a unique time in the botanical collecting record for this taxon in this state (additive with the previous two scales; Fig. 3). We observed (and our models predicted) that the additive unique county and locality probabilities were always less than 0.5 for both herbarium size classes and all four species statuses, indicating that a specimen has a probability of providing a unique record at the temporal scale more than half of the time. Therefore, specimens in herbaria often represent repeated

TABLE 1. Number of unique collecting events represented by unduplicated and duplicated specimens held in large vs. small herbaria.

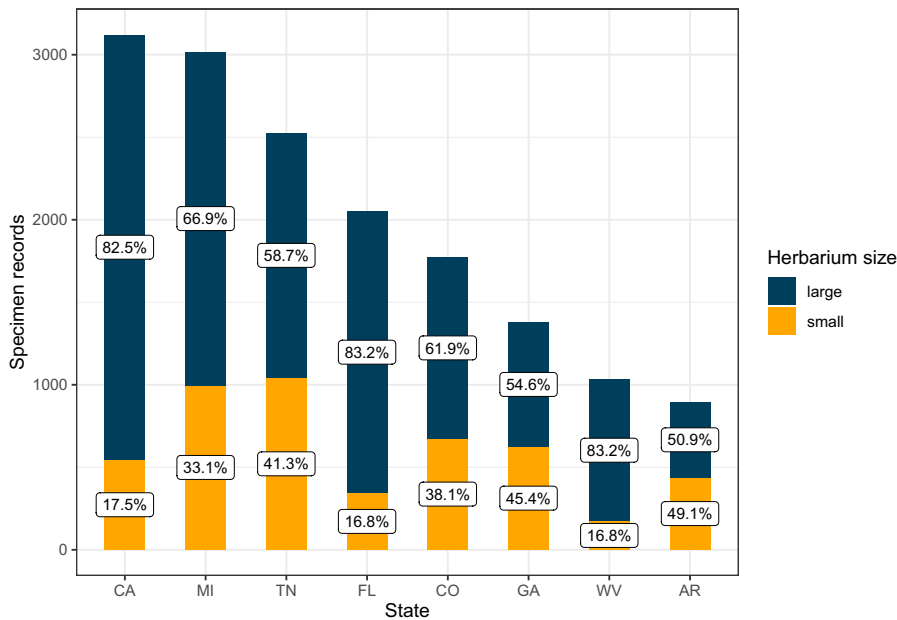| Duplicate type	Large herbaria	Small herbaria
Unduplicated specimens	9415 (83%)	4456 (89%)
Duplicated specimens held only by large herbaria	1635 (14.4%)	N/A
Duplicated specimens held only by small herbaria	N/A	423 (8.4%)
Duplicated specimens held by large and small herbaria	289 (2.6%)	130 (2.6%)
Total unique collecting events	**11,339 (100%)**	**5009 (100%)**

FIGURE 1. Number of specimen records included in this study's primary analysis data set that were contributed by large (≥100,000 specimens) and small herbaria (<100,000 specimens) in each participating state.

collections from the same localities over time, possibly due to habitat loss, proximity to the herbarium, other access-related factors such as permits for collecting, or an emphasis on known botanical areas of interest.

We suspect that small herbaria may be especially relevant to research focused on regionally occurring taxa, as evidenced by the 17-percentage-point increase between the total number of specimens held by small (vs. large) herbaria contacted for this project (12.5%), and the number of relevant specimens that these small herbaria contributed to the project data set (29.7%), which had a focus on regional taxa. Small herbaria likely have staff and students focused on collecting specimens from their own local vicinity. Moreover, student collections may be repeated over time from the same localities due to the nature of course assignments or access to certain sites known by the curator of the herbarium. For a complete understanding of species distributions, a thorough sampling of collections of all sizes is warranted, and based on the idiosyncratic nature of collections and curatorial research interest, it is difficult to predict a priori which herbaria might be excluded without resulting data loss.

While a thorough sampling of many herbaria is challenging in person, digitization offers an excellent compromise. We recommend including herbaria of all sizes equally in digitization efforts and encouraging the mobilization of digitized data and media to biodiversity data aggregators such as iDigBio (www.idigbio.org) and the Global Biodiversity Information Facility (www.gbif.org). Our data collection was complicated by the uneven distribution of digitally accessible data across herbaria. For collections that already had a significant amount of data digitized and available online, e.g., on the Consortium of California Herbaria portal, we downloaded those data directly, whereas for collections without an online presence of specimen records, we requested data from each herbarium. If data from portals were present, but not complete, then we missed some existing data because we did not contact individual herbaria if data

for our target taxa were present in an online format. Paradoxically, it is therefore possible that we received more complete data from herbaria without a digital presence at the time of data collection. Our own experience highlights the importance of improving digital accessibility for all herbarium specimens.

In the last decade, there has been a genuine effort to include small collections in digitization projects funded through the National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) program. The SouthEast Regional Network of Expertise and Collections (SERNEC; sernecportal.org) and the Southern Rockies projects (Allen, 2018) are two examples of how small herbaria have successfully been integrated in digitization projects beyond the scope of what they might have had the individual capacity to do otherwise. We contend that continuing to digitize herbaria of all sizes will ameliorate the lack-of-data situation to some degree, but we also realize that continued regional collecting is necessary. Prather et al. (2004) found that local collecting is on the decline in two-thirds of the herbaria surveyed, regardless of herbarium size. Ferro and Flick (2015) discovered that bias in entomology collections has a serious effect on species distribution modelling and that the number of collections contributing specimens, rather than the number of localities sampled or specimens themselves, is a better indicator of exhaustiveness in avoiding bias. They also argued that "maintenance and growth of numerous, regional natural history collections is important" (Ferro and Flick, 2015, p. 424), which applies to herbaria as it does to entomology collections. Not only do staff at small herbaria curate and make specimens accessible, but they also foster regional expertise that may not be accurately captured in specimen data alone. For instance, historic collecting localities can be notoriously difficult to interpret for modern georeferencing, and even more recently collected specimens may use local place names to describe localities. Collections with a regional focus tend to be associated with people who are more familiar with the surrounding geography and to whom local place names are meaningful. This regional knowledge translates into georeferencing precision and accuracy, which are the most highly desirable qualities sought by users working with species occurrence data (Ariño et al., 2013).

Digitization, continued collecting, and maintaining and enhancing regional biogeographical knowledge require the recognition of herbaria as critical research infrastructure and the understanding that in the USA this infrastructure comprises 686 individual herbaria, 85% of which are small collections with fewer than 100,000 specimens (Thiers and Ramirez, 2020). Our herbaria of all sizes continue to need significant financial support, and to this extent, it is key for university administrators to understand the value of natural history collections. We provide a template letter of advocacy from an herbarium curator to an institutional administrator to assist in starting this discussion for readers in a position to do so (Appendix S6).
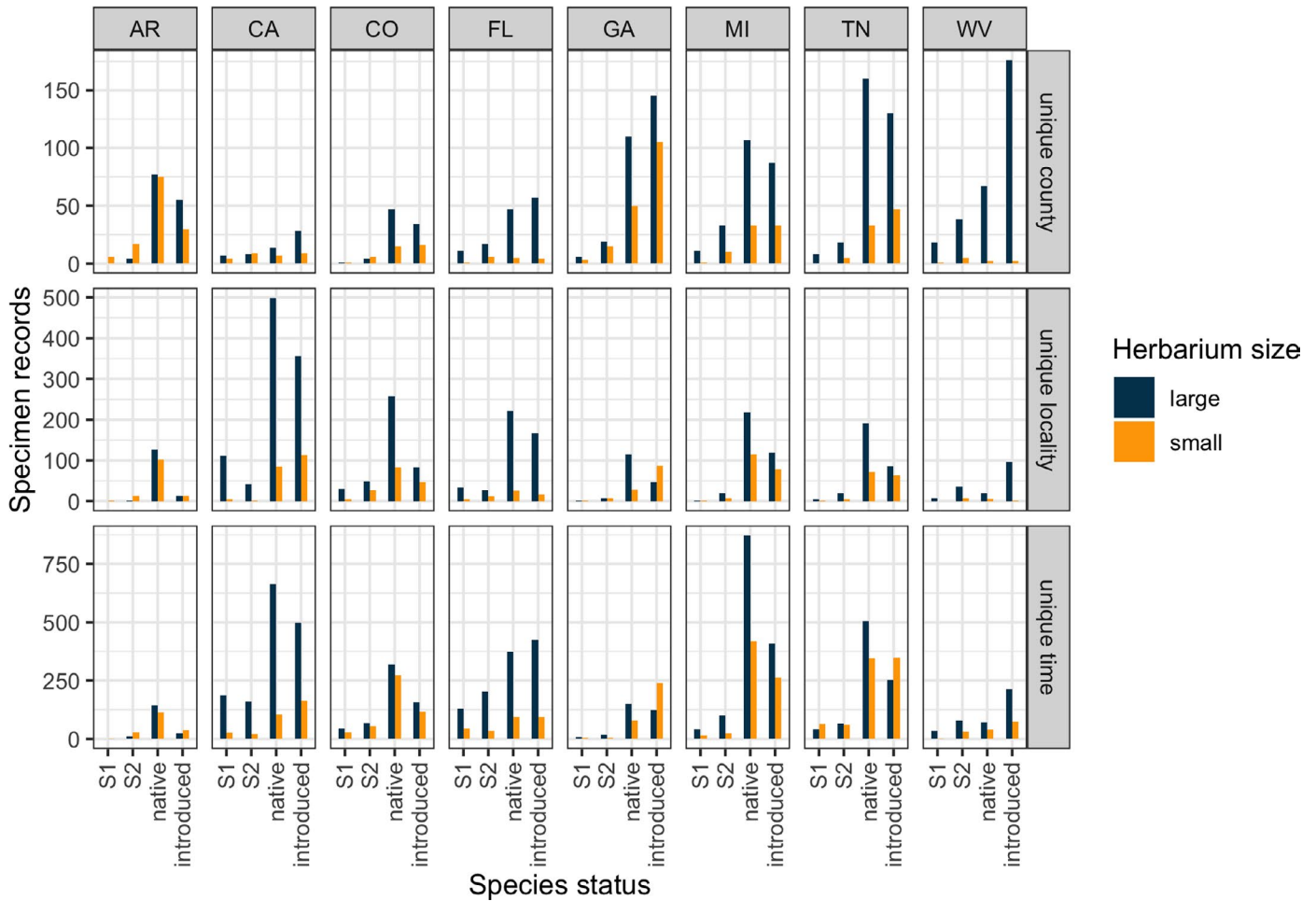
FIGURE 2. Number of specimen records included in this study's primary analysis data set that were contributed by large (≥100,000 specimens) and small herbaria (<100,000 specimens) in each participating state, faceted by scale of uniqueness (county, locality, temporal) and species status category (S1, S2, common native, introduced).

Better access to digitized specimen data will allow future studies to address the contributions of out-of-state herbaria to in-state biogeographical knowledge, which this study did not. We decided not to include specimens held in out-of-state herbaria in our analyses because of the complexity involved in data gathering, although we think that doing so would affect our narrative in regard to duplicate specimens and specimen uniqueness. Out-of-state holdings can contain critical specimens for our understanding of certain areas. For example, field research for the Flora of the Four Corners Region (Heil et al., 2013) resulted in a large number of collections from four states since the flora followed an ecological rather than a political boundary. Most of the specimens were deposited in the San Juan College Herbarium (SJNM; Farmington, New Mexico, USA), since the principal author curates the herbarium there. Another example of important Colorado specimens being held out-of-state comes from the large floristic inventory program of the Rocky Mountain Herbarium (RM) at the University of Wyoming, Laramie, Wyoming. This program was initiated in 1978 and resulted in more than 60 floristic studies across 13 states, contributing more than 640,000 specimens total and over 107,000 specimens from Colorado (Rocky Mountain Herbarium, 2020). Moreover, we know that specimen collecting and duplicate sharing can be influenced by proximity and social connections rather than the confines of a state's boundaries.

For example, in Arkansas, multiple small herbaria shared duplicates with the nearby but out-of-state herbarium at the University of Louisiana at Monroe (NLU; Monroe, Louisiana, USA), a large herbarium that makes a particularly interesting example because it was orphaned by the university and subsequently transferred to the Botanical Research Institute of Texas (BRIT) in 2017. A future avenue for research aimed at understanding knowledge gaps in biogeographic patterns from existing data should investigate specimen contributions held uniquely outside the state boundaries from where the specimens were collected.

CONCLUSIONS

In sum, herbaria of all sizes are important resources for preserving and expanding our knowledge of phytogeography. Small herbaria are crucial components of this research infrastructure because they contain records that fill gaps (this study), because more collections ameliorate bias (Soberon, 1999; Ferro and Flick, 2015; Krishtalka and Humphrey, 2000), and because most herbaria in the USA are small (Thiers, 2020; Thiers and Ramirez, 2020). Digitization and data sharing have removed the historical logistical barrier for a researcher having to visit many separate collections to assess specimen
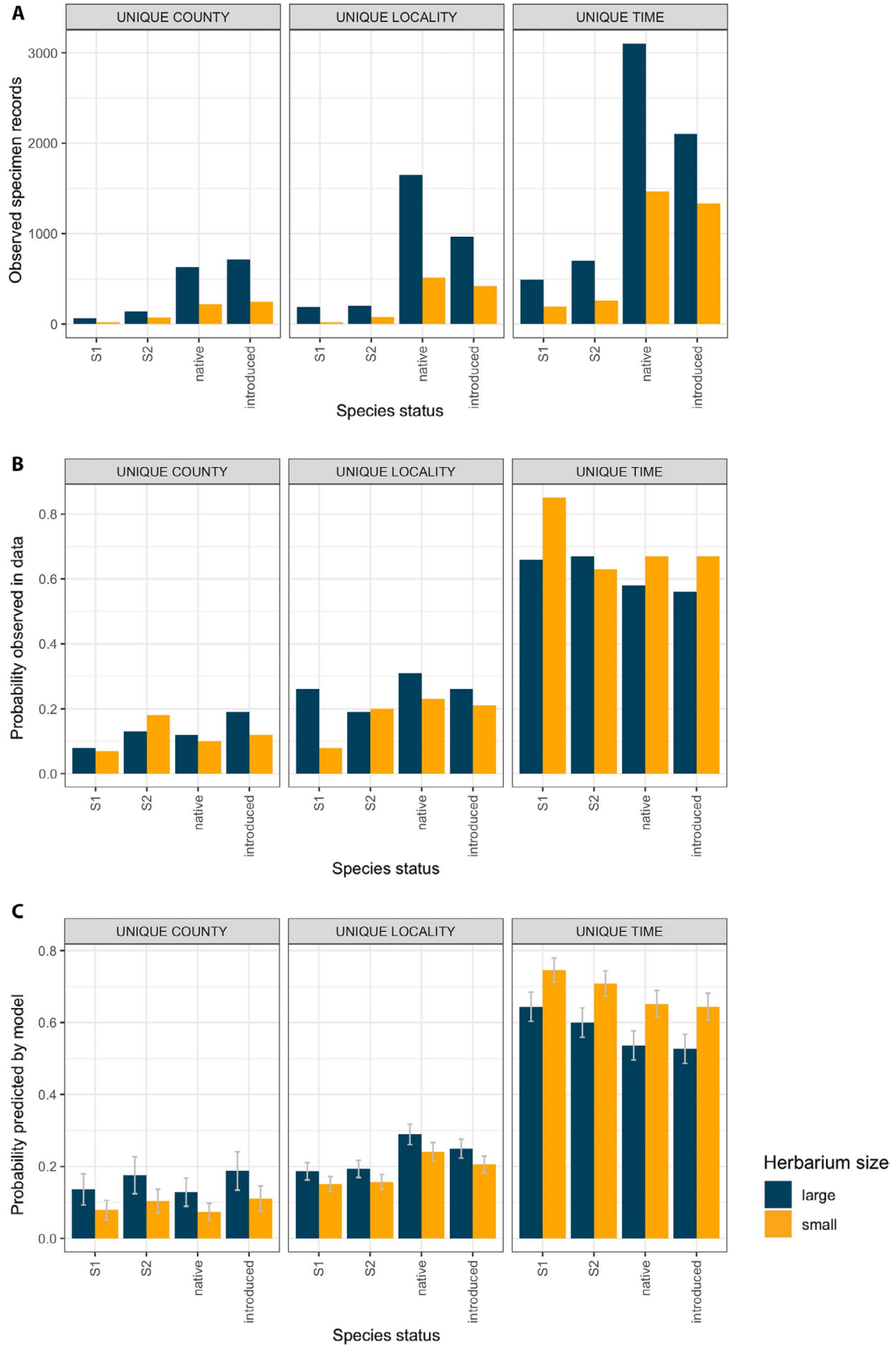
FIGURE 3. Assessment of model validity in predicting the probability that a specimen represents unique information at different biogeographic scales by comparing (A) observed specimen records and (B) probability in observed data to (C) probability predicted by model. Given that the herbarium size class and species status of a specimen are inherent attributes of the specimen, this figure illustrates the scale of biogeographic uniqueness at which a particular specimen might be expected to contribute.

TABLE 2. Model selection results of specimen uniqueness at the county, locality, and temporal scales. Shown are the degrees of freedom, Akaike information criterion (AIC) values, ΔAIC values, and AIC weights. In each model, state is included as a random variable.

Response variable	Model predictors	df	AIC	ΔAIC	AIC weight
County scale uniqueness	Size class + species status	5	10958	0	1
	Size class	2	11022	64.4	0
	Species status	4	11076	118.3	0
	No predictor	1	11123	165.5	0
Locality scale uniqueness	Size class + species status	5	17332	0	1
	Species status	4	17365	32.7	0
	Size class	2	17422	90.0	0
	No predictor	1	17459	126.9	0
Temporal scale uniqueness	Size class + species status	5	20408	0	1
	Size class	2	20460	52.8	0
	Species status	4	20566	158.5	0
	No predictor	1	20616	208.8	0

holdings or acquire digital data, so digital data sharing is an essential strategy for democratizing access to all herbaria.

AUTHOR CONTRIBUTIONS

T.D.M., E.R.K., J.R.C., E.L.G., P.D.L., R.M., A.B.M., G.N., M.S., and A.K.M. conceived of the idea and gathered data from their state herbaria. Countless hours were spent on conference calls to strategize and implement a uniform approach to gathering and collating data. A.K.M. provided initial leadership and momentum. E.R.K. and T.D.M. georeferenced any specimens for which it was necessary. E.R.K. conducted the data compilation and preliminary analyses. D.L.S. and E.R.K. conducted the modeling analyses. T.D.M. and E.R.K. led the writing of the manuscript. All authors contributed to and edited the manuscript.

DATA AVAILABILITY

Data collated for the purposes of this study and the analysis code written in R are archived and available on Zenodo at https://doi.org/10.5281/zenodo.3937865. A version of the analysis code rendered for viewing in a web browser can be found at https://ekrimmel.github.io/marsico-et-al-2020/Marsico-et-al-2020_v4.html.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Excel spreadsheet documenting all 320 taxa used for this project (8 states × 40 taxa per state), including data sources for each species category.

APPENDIX S2. Excel spreadsheet documenting all herbaria contacted to provide data for this project, including information on collection size and data contribution.

APPENDIX S3. Excel spreadsheet providing a data dictionary for fields in our data and details about any transformations done to them during compilation.

APPENDIX S4. Excel spreadsheet of results from analysis to determine unique specimen records contributed to this study by large (≥100,000 specimens) and small herbaria (<100,000 specimens) in each participating state, faceted by scale of uniqueness (county, locality, temporal) and species category (S1, S2, common native, introduced). Figure 2 is a visualization of these data.

APPENDIX S5. Analysis summary (equivalent to Appendix S4), duplicate summary (equivalent to Table 1), and modelling results (equivalent to Table 2) from an alternative analysis of data using cutoff of 175,000 specimens to differentiate between large and small herbaria.

APPENDIX S6. Example letter to university/institution administrators highlighting the work in this paper so that curators can help justify the research contributions made by small herbaria.

LITERATURE CITED

Allen, J. R. 2018. Advancing digitization in the southern Rocky Mountain region. *The Vasculum* (newsletter of the Society of Herbarium Curators) 13: 9–12.

Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* 66: 912–918.

Ariño, A. H., V. Chavan, and D. P. Faith. 2013. Assessment of user needs of primary biodiversity data: Analysis, concerns, and challenges. *Biodiversity Informatics* 8: 59–63.

Ball-Damerow, J. E., L. Brenskelle, N. Barve, P. S. Soltis, P. Sierwald, R. Bieler, R. LaFrance, et al. 2019. Research applications of primary biodiversity databases in the digital age. *PLoS One* 149: e0215794.

Barkworth, M., and Z. Murrell. 2012. The US Virtual Herbarium: working with individual herbaria to build a national resource. *ZooKeys* 209: 55–73.

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.

Bivand, R., T. Keitt, and B. Rowlingson. 2019. rgdal: Bindings for the Geospatial Data Abstraction Library (GDAL). R package version 1.4-7. Website: https://CRAN.R-project.org/package=rgdal [accessed 01 March 2019].

Bivand, R., and N. Lewin-Koh. 2019. maptools: Tools for handling spatial objects. R package version 0.9-8. Website: https://CRAN.R-project.org/package=maptools [accessed 01 March 2019].

Bivand, R., and C. Rundel. 2019. rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.5-2. Website: https://CRAN.R-project.org/package=rgeos [accessed 01 March 2019].

Bloom, T. D. S., A. Flower, and E. G. DeChaine. 2018. Why georeferencing matters: introducing a practical protocol to prepare species occurrence records for spatial analysis. *Ecology and Evolution* 8: 765–777.

Brenskelle, L., B. J. Stucky, J. Deck, R. Walls, and R. P. Guralnick. 2019. Integrating herbarium specimen observations into global phenology data systems. *Applications in Plant Sciences* 7: e01231.

Calinger, K. M., S. Queenborough, and P. S. Curtis. 2013. Herbarium specimens reveal the footprint of climate change on flowering trends across north-central North America. *Ecology Letters* 16: 1037–1044.

Casas-Marce, M., E. Revilla, M. Fernandes, A. Rodríguez, M. Delibes, and J. A. Godoy. 2012. The value of hidden scientific resources: preserved animal specimens from private collections and small museums. *BioScience* 62: 1077–1082.

Cobb, N. S., L. F. Gall, J. M. Zaspel, N. J. Dowdy, L. M. McCabe, and A. Y. Kawahara. 2019. Assessment of North American arthropod collections: Prospects and challenges for addressing biodiversity research. *PeerJ* 7: e8086.

Daru, B. H., D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. S. Whitfeld, T. G. Seidler, et al. 2017. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* 217: 939–955.

Davis, C. C., C. G. Willis, B. Connolly, C. Kelly, and A. M. Ellison. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany* 102: 1599–1609.

Feeley, K. J., and M. R. Silman. 2011. Keep collecting: Accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions* 1–9: 1132–1140.

Ferro, M. L., and A. J. Flick. 2015. Collection bias and the importance of natural history collections in species habitat modeling: a case study using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), a critique of gbif.org. *Coleopterists Bulletin* 69: 415–425.

Fox, J., and S. Weisberg. 2019. car: companion to applied regression. R package version 3.0-5. Website: https://cran.r-project.org/web/packages/car/index.html [accessed 01 June 2019].

Gehan, M. A., and E. A. Kellogg. 2017. High-throughput phenotyping. *American Journal of Botany* 104: 505–508.

Glon, H. E., B. W. Heumann, J. R. Carter, J. M. Bartek, and A. K. Monfils. 2017. The contribution of small collections to species distribution modelling: A case study from Fuireneae (Cyperaceae). *Ecological Informatics* 42: 67–78.

Gropp, R. E. 2003. Are university natural science collections going extinct? *BioScience* 53: 550.

Guralnick, R., and A. Hill. 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* 25: 421–428.

Heberling, J. M., and B. L. Isaac. 2017. Herbarium specimens as exaptations: new uses for old collections. *American Journal of Botany* 104: 963–965.

Heberling, J. M., L. A. Prather, and S. J. Tonsor. 2019. The changing uses of herbarium data in an era of global change: an overview using automated content analysis. *BioScience* 69: 812–822.

Heil, K. D., S. L. O'Kane, L. M. Reeves, and A. Clifford. 2013. Flora of the Four Corners Region: vascular plants of the San Juan River Drainage, Arizona, Colorado, New Mexico, and Utah. Monographs in Systematic Botany from the Missouri Botanical Garden, vol. 124. Missouri Botanical Garden Press, St. Louis, MO, USA.

Krishtalka, L., and P. S. Humphrey. 2000. Can natural history museums capture the future? *BioScience* 50: 611–617.

Lavoie, C. 2013. Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics* 15: 68–76.

Lendemer, J., B. Thiers, A. K. Monfils, J. Zaspel, E. R. Ellwood, A. Bentley, K. Levan, et al. 2020. The Extended Specimen Network: a strategy to enhance US biodiversity collections, promote research and education. *BioScience* 70: 23–30.

Lenth, R. 2020. emmeans: estimated marginal means, aka least-squares means. Website: https://CRAN.R-project.org/package=emmeans [accessed 06 May 2020].

López, A., and A. B. Sassone. 2019. The uses of herbaria in botanical research. A review based on evidence from Argentina. *Frontiers in Plant Science* 10: 1363.

Lughadha, E. N., B. E. Walker, C. Canteiro, H. Chadburn, A. P. Davis, S. Hargreaves, E. J. Lucas, et al. 2018. The use and misuse of herbarium specimens in evaluating plant extinction risks. *Philosophical Transactions of the Royal Society, B, Biological Sciences* 374: 20170402.

McCartha, G. L., C. M. Taylor, A. van der Ent, G. Eschevarria, D. M. Navarrete Gutiérrez, and A. J. Pollard. 2019. Phylogenetic and geographic distribution of nickel hyperaccumulation in neotropical *Psychotria*. *American Journal of Botany* 106: 1377–1385.

Meyer, C., P. Weigelt, and H. Kreft. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19: 992–1006.

Miller-Rushing, A. J., R. B. Primack, D. Primack, and S. Mukunda. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. *American Journal of Botany* 93: 1667–1674.

Monfils, A. K., E. R. Krimmel, J. M. Bates, J. E. Bauer, M. W. Belitz, B. C. Cahill, A. M. Caywood, et al. 2020. Regional collections are an essential component of biodiversity research infrastructure. *BioScience* biaa102.

Navarro, D. J. 2015. lsr: companion to "Learning statistics with R". R package version 0.5. Website: https://CRAN.R-project.org/package=lsr [accessed 01 June 2019].

O'Connell, A. F., A. T. Gilbert, and J. S. Hatfield. 2004. Contribution of natural history collection data to biodiversity assessment in national parks. *Conservation Biology* 18: 1254–1261.

OpenRefine Core Team. 2018. OpenRefine: a free, open source power tool for working with messy data and improving it, version 2.8 for Mac. Website: https://www.openrefine.org/ [accessed 01 January 2018].

Page, L. M., B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* 65: 841–842.

Park, I. W., and M. D. Schwartz. 2015. Long-term herbarium records reveal temperature-dependent changes in flowering phenology in the southeastern USA. *International Journal of Biometeorology* 59: 347–355.

Pearse, W. D., C. C. Davis, D. W. Inouye, R. B. Primack, and T. J. Davies. 2017. A statistical estimator for determining the limits of contemporary and historic phenology. *Nature Ecology & Evolution* 1: 1876–1882.

Pearson, K. D. 2019. Spring- and fall-flowering species show diverging phenological responses to climate in the southeast USA. *International Journal of Biometeorology* 63: 481–492.

Prather, L. A., O. Alvarez-Fuentes, M. H. Mayfield, and C. J. Ferguson. 2004. The decline of plant collecting in the United States: A threat to the infrastructure of biodiversity studies. *Systematic Botany* 29: 15–28.

Pyke, G. H., and P. R. Ehrlich. 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological Reviews* 85: 247–266.

QGIS Development Team. 2019. QGIS Geographic Information System, version 3.4 for Mac. Website: http://qgis.osgeo.org [accessed 01 January 2019].

R Core Team. 2019. R: A language and environment for statistical computing, version 3.6.1 for Mac. Website: https://www.R-project.org/ [accessed 05 July 2019].

Rawal, D. S., S. Kasel, M. R. Keatley, and C. R. Nitschke. 2015. Herbarium records identify sensitivity of flowering phenology of eucalypts to climate: implications for species response to climate change. *Austral Ecology* 40: 117–125.

Rios, N. 2019. GEOLocate software for georeferencing natural history data. Website: http://www.geo-locate.org [accessed 01 January 2019 through 30 June 2019].

Ristaino, J. B., C. T. Groves, and G. R. Parra. 2001. PCR amplification of the Irish potato famine pathogen from historic specimens. *Nature* 411: 695–697.

Rocky Mountain Herbarium. 2020. Projects by graduate students in floristics/staff/associates. University of Wyoming, Laramie, WY, USA. Website: https://www.uwyo.edu/botany/rocky-mountain-herbarium/study-areas.pdf.

Snow, N. 2005. Successfully curating smaller herbaria and natural history collections in academic settings. *BioScience* 55: 771–779.

Soberon, J. 1999. Linking biodiversity information sources. *Trends in Ecology & Evolution* 14: 291.

Soltis, P. S. 2017. Digitization of herbaria enables novel research. *American Journal of Botany* 104: 1281–1284.

Sorrie, B., and A. Weakley. 2001. Coastal plain vascular plant endemics: Phytogeographic patterns. *Castanea* 66(1/2): 50–82.

Thiers, B. M. 2020. The world's herbaria 2019: A summary report based on data from *Index Herbariorum*. Website: http://sweetgum.nybg.org/science/docs/The_Worlds_Herbaria_2019.pdf.

Thiers, B. M., and J. Ramirez. 2020. Index Herbariorum API, version 1.0. Website: http://sweetgum.nybg.org/science/api/v1/institutions/search?country=u.s.a.&download=yes [accessed 13 February 2020].

Ward, D. F. 2012. More than just records: Analysing natural history collections for biodiversity planning. *PLoS One* 7: e50346.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. François, G. Grolemund, et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4: 1686.

Willis, C. G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, et al. 2017. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution* 32: 531–546.

Winker, K. 2004. Natural history museums in a postbiodiversity era. *BioScience* 54: 455–459.

Zhu, Hao. 2019. kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. Website: https://CRAN.R-project.org/package=kableExtra [accessed 01 June 2019].