Ecological Informatics 42 (2017) 67–78

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

The contribution of small collections to species distribution modelling: A case study from Fuireneae (Cyperaceae)

Heather E. Glon[a,b,*], Benjamin W. Heumann[a,c], J. Richard Carter[d], Jessica M. Bartek[d], Anna K. Monfils[a,b]

[a] Central Michigan University, Institute for Great Lakes Research, Mount Pleasant, MI, USA
[b] Central Michigan University, Department of Biology, Mount Pleasant, MI, USA
[c] Central Michigan University, Department of Geography, Center for Geographic Information Science, Mount Pleasant, MI, USA
[d] Department of Biology, Valdosta State University, Valdosta, GA, USA

ARTICLE INFO

Keywords:
Species distribution modelling
Global biodiversity information facility
Natural history collections
MaxEnt

ABSTRACT

The recent and rapid digitization of biodiversity data from natural history collection (NHC) archives has enriched collections based data repositories; this data continues to inform studies of species' geographic distributions. Here we investigate the relative impact of plant data from small natural history collections (collections with < 100,000 specimens) on species distributional models in an effort to document the potential of data from small NHCs to contribute to and inform biodiversity research. We modelled suitable habitat of five test case species from Fuireneae (Cyperaceae) in the United States using specimen records available via the Global Biodiversity Information Facility and that of data ready to mobilize from two regional small herbaria. Data were partitioned into three datasets based on their source: 1) collections-based records from large NHCs accessed GBIF, 2) collections-based records from small NHCs accessed from GBIF, and 3) collections-based records from two small regional herbaria not yet mobilized to GBIF. We extracted and evaluated the ecological niche represented for each of the three datasets by applying dataset occurrences to 14 environmental factors, and we modelled habitat suitability using Maxent to compare the represented distribution of the environmental values among the datasets. Our analyses indicate that the data from small NHCs contributed unique information in both geographic and environmental space. When data from small collections were combined with data from large collections, species models of the ecological niche resulted in more refined predictions of habitat suitability, indicating that small collections can contribute unique occurrence data which enhance species distribution models by bridging geographic collection gaps and shifting modelled predictions of suitable habitat. Inclusion of specimen records from small collections in ongoing digitization efforts is essential for generating informed models of a species' niche and distribution.

1. Introduction

1.1. Background

Natural history collections (NHCs) preserve and archive biological specimens with their associated occurrence and locality data. These data document species diversity over time and across the globe, and are being used to address scientific issues of global concern including climate change, infectious diseases spread, invasive species distributions, habitat and biodiversity loss, and conservation of natural resources (see Chapman and Speers, 2005; Crawford and Hoagland, 2009; Davis et al., 2015; Faith et al., 2013; Gallagher et al., 2009; Lavoie, 2013; Newbold, 2010; Pyke and Ehrlich, 2010; Robbirt et al., 2011; Suarez and Tsutsui, 2004; Wen et al., 2015). The United States (US) collections community has embraced a national digitization effort with the goal to image all US specimens, transcribe associated collections based data, and mobilize digitized records into a common portal over a 10-year time period (Beach et al., 2012). This massive coordinated digitization effort will standardize data delivery, and provide unrestricted and centralized access to valuable and informative specimen-based biodiversity data (Beaman and Cellinese, 2012; Gaiji et al., 2013).

The national digitization of biological specimens and their associated data is inclusive of all research biological and paleontological collection types, sizes, and taxonomic groups. As the digitization initiative was developing, the Network for Integrated Biocollections (NIBA) Implementation plan specifically addressed the importance of

* Corresponding author at: Central Michigan University, Mount Pleasant, MI, USA.
E-mail address: dame1h@cmich.edu (H.E. Glon).

http://dx.doi.org/10.1016/j.ecoinf.2017.09.009
Received 10 May 2017; Received in revised form 26 September 2017; Accepted 27 September 2017
Available online 30 September 2017

including digitization of specimens from small NHCs (defined as NHCs with < 100,000 specimens per Monfils and Nelson, 2014). Small NHCs are typically regional in scope with a defined ecological, taxonomic or geographic focus (i.e. small colleges and universities, biological stations, field stations). Often, data from small collections are not included in literature reviews of the regional flora and fauna, so the diversity they archive is not represented in the collective records of species inventories, field guides, local flora, and monographic studies. Increasing the availability of small collections in aggregated databases (i.e. GBIF, VERTNET, iDigBio Portal, SEINet) has the potential to offset spatial sampling bias inherently present in NHC data (Chauvel et al., 2006; Ferro and Flick, 2015). Despite such indications of their value, small herbaria are under-consulted for scientific purposes when compared to large herbaria (Lavoie, 2013). Using the data from *Index Herbariorum* (Theirs, 2014), over 83% of herbaria are classified as small NHCs and preliminary work in the mammal community indicates similar numbers relative to other types of NHCs (M. Revelez, SPNHC 2015). This directive to include small collection data has been championed by the Advancing Digitization of Biodiversity Collections (ADBC) NSF program, the Integrated Digitized Biocollections HUB (iDigBio; http://www.idigbio.org), and the Biodiversity Collections Network (BCoN; http://www.bcon.aibs.org). The result, a Small Collections Network (SCNet; http://smallcollections.net), has been successful in raising the profile and providing resources and networking capabilities to small NHCs (Monfils and Nelson, 2014).

The specimen-based data entered into the Global Biodiversity Information Facility (GBIF) from natural history collections have been used extensively in species distribution modelling studies (Gaiji et al., 2013). The reliability of collections based data for species richness and distribution modelling has come into question with specific concerns regarding presence only data and potential collection bias (see Beck et al., 2013, 2014; Davis et al., 2015; Ferro and Flick, 2015; García-roselló et al., 2014; Graham et al., 2004; Newbold, 2010; Yesson et al., 2007). Bias in sampling that can detrimentally influence species distribution models encompasses mainly spatial scales, but also can include taxonomic, environmental, and temporal scales (Ferro and Flick, 2015; Graham et al., 2004; Guillera-Arroita et al., 2015; Meyer et al., 2016; Newbold, 2010). Spatial bias occurs due to uneven sampling across a geographic space or incorrect georeferencing, which has the potential of environmentally biased occurrence data depending on the species and the dataset (Hortal et al., 2008; Newbold, 2010; Newbold et al., 2009). Natural history collections are susceptible to biases resulting from intensive collecting in specific geographical areas or of rare taxa (Lavoie, 2013), which can skew model results, creating models with inaccurate portrayals of biodiversity (Austin, 2007). The most reliable species distribution models employ data that fully characterize a species' range of environmental conditions.

1.2. Objectives

In this study, we examine how specimen based data from small herbaria can inform geographic species distribution modelling. Taxa from tribe Fuireneae (Cyperaceae; sedges) were used as a case study. Collections based data was sourced from The Global Biodiversity Information Facility (GBIF; http://www.GBIF.org); currently the largest aggregator of global biodiversity collections based records (Berendsohn et al., 2010; Roberts et al., 2015). Additional data, ready for mobilization but not yet publicly available in GBIF at the time, was sourced from Valdosta State (VSC) and the Central Michigan University (CMC) Herbaria. We used three data sets in our analysis: GBIF data from large collections (defined as natural history collections > 100,000 specimens), all GBIF data (from large and small collections), and all GBIF data combined with data from VSC and CMC. These data were applied to Species Distribution Modelling (SDM) analyses using Maxent to 1) determine if data from small herbaria increases our ability to capture the environmental range within a species, and 2) investigate how data from small herbaria contributes to the predictive maps for habitat suitability.

2. Materials and methods

2.1. Taxa and study area

Fuireneae (Cyperaceae) is largely composed of obligate wetland species, including narrow endemics, with restricted habitat types and specific environmental requirements. Four of the six Fuireneae genera used in this study occur in the contiguous United States: *Bolboschoenus* (Asch) Palla (5 of the worldwide species (15) and sub-species (2) occur in the contiguous United States), *Fuirena* Rottb. (8 of the worldwide species (58) and sub-species (5) occur in the contiguous United States), *Schoenoplectiella* Lye (9 of the worldwide species (52) and sub-species (5) occur in the contiguous United States), and *Schoenoplectus* (Rchb.) Palla (14 of the worldwide species (29) and sub-species (7) occur in the contiguous United States; species counts from Kew World Checklist online and Flora of North America; Flora of North America Editorial Committee, 2014; Govaerts et al., 2014). The study area was defined within the continental United States based on the availability of consistent and high quality data from both environmental and collections databases. In the United States, taxa within the tribe are both widespread (e.g. *Schoenoplectus pungens*) and narrow endemics (e.g. *Schoenoplectiella purshiana*). Additionally, the United States contains several federally or state listed Fuireneae species (e.g. *Fuirena squarrosa*) and is a site of rapid radiation for *Schoenoplectus* (Shiels et al., 2014).

2.2. Species occurrence datasets

Records were downloaded for all species in Fuireneae from GBIF (accessed January 15th, 2015). We included additional data from Central Michigan University (CMC; 22,000 specimens) and Valdosta State University (VSC; 71,000 specimens). These herbaria were not mobilized to GBIF at the time of data compilation but are currently available online in the GBIF portal. Three subsets of data were defined: GBIF specimen-based data from large herbaria (GBIF Large); GBIF specimen-based data from small herbaria (GBIF Small); and regional specimen-based data not mobilized to GBIF (CMC/VSC).

Species names and occurrence data were cleaned for use in modelling (Fig. 1). Records were limited to data from preserved specimens in the contiguous United States and names were accepted to the species rank. Names were reconciled with current accepted taxonomy using the Kew World Checklist (http://apps.kew.org/wcsp/), Tropicos (http://www.tropicos.org), and the International Plant Names Index (http://www.ipni.org). Environmental variables limited the resolution of georeferenced occurrence data to 5′ (10 km^2 at the equator); any records with insufficient geographic precision (i.e. those with 0.01 decimal degrees were removed). Records in the CMC and VSC dataset without associated geographical coordinates were georeferenced using GEOLocate at the respective herbaria (Rios and Bart, 2010). Records with TRS (Township, Range and Section) data were georeferenced using the Bureau of Land Management (BLM) batch processing or by single point translation using Earthpoint (http://www.earthpoint.us/Townships.aspx). Specimens that could not be georeferenced reliably or georeferenced only to county centroid were removed from the dataset. As Maxent uses only one occurrence record per grid cell to assist in reducing sampling bias, grid replicate occurrences from each dataset

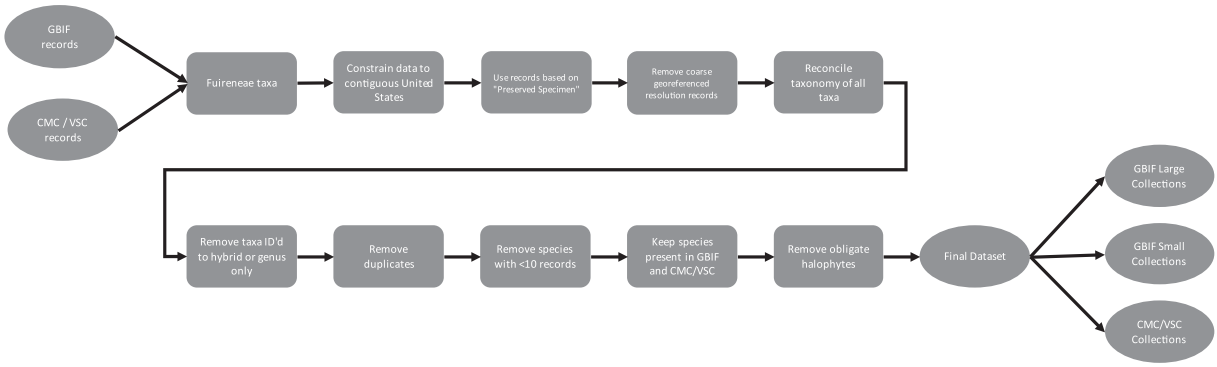

Fig. 1. Flowchart of the pre-processing of species occurrence data from GBIF and CMC/VSC collections included in analyses. Ovals represent input and output datasets. Boxes represent filters imposed upon the input datasets to reduce the initial datasets to the final five species. The final data was split into three separate datasets: large collections data from GBIF (GBIF Large collection), small collections data from GBIF (GBIF Small collection), and data from the CMC/VSC collections (CMC/VSC collection). The original number of downloaded records in GBIF was 15,967; this dataset was split into two subsets of 1269 (GBIF Large) and 122 (GBIF Small) records. The CMC/VSC dataset was reduced from 538 to 127 records.

were removed using ENMTools (Warren et al., 2010). Further trimming of species occurred due to a lack of available suitable environmental data (e.g. obligate halophytic species), or contributing fewer than 25 or 10 occurrences from large or small collections, respectively (Breiner et al., 2015; Pearson et al., 2007; Wisz et al., 2008). Because this analysis is less sensitive to low sample size, datasets containing a smaller amount of occurrence points (≥ 25) were allowed.

2.3. Environmental factors

We examined 19 bioclimatic environmental factors available from the WorldClim database (Hijmans et al., 2005). We used a spatial resolution of 5 arcmin and clipped the layers in ArcMap to the extent of the contiguous United States. As model performance can be influenced by inclusion of too many inter-correlated environmental factors, we used ENMTools (Warren et al., 2008) to generate Pearson's correlation coefficients for pairwise comparisons between all WorldClim variables (Appendix A, Table A.1–A.2; Heikkinen et al., 2006; Guo, 2013). Of these bioclimatic factors, 12 of 19 were removed due to high correlation ($r > |0.7|$; Dormann et al., 2012), leaving seven factors (Table 1).

Soil data was downloaded from the Digital General Soil Map of the United States (STATSGO2), and reprocessed as percent abundance of the 12 soil orders across the United States in each cell. These layers were resampled in ArcMap to the same processing extent and resolution as the WorldClim bioclimatic layers. The 5 soil orders that rarely or never co-occurred with the species data were removed (STATSGO2, Soil Survey Staff, n.d.; Table 1).

Table 1
All WorldClim Bioclimatic and StatsGo2 soil variables. Environmental factors in bold were included in the Maxent model for the niche modelling of Fuireneae. All other factors were not included on the basis of being highly correlated or not relevant to the study.

Factor	Measures	Source
BIO1	Annual Mean Temperature	WorldClim
BIO2	**Mean Diurnal Range (°C, Mean of monthly temp, max-min)**	**WorldClim**
BIO3	Isothermality (BIO2/BIO7)(*100)	WorldClim
BIO4	**Temperature Seasonality (°C, standard deviation *100)**	**WorldClim**
BIO5	**Max Temperature of Warmest Month (°C)**	**WorldClim**
BIO6	Min Temperature of Coldest Month	WorldClim
BIO7	Temperature Annual Range (BIO5-BIO6)	WorldClim
BIO8	**Mean Temperature of Wettest Quarter (°C)**	**WorldClim**
BIO9	Mean Temperature of Driest Quarter	WorldClim
BIO10	Mean Temperature of Warmest Quarter	WorldClim
BIO11	Mean Temperature of Coldest Quarter	WorldClim
BIO12	Annual Precipitation	WorldClim
BIO13	Precipitation of Wettest Month	WorldClim
BIO14	Precipitation of Driest Month	WorldClim
BIO15	**Precipitation Seasonality (mm, coefficient of variation)**	**WorldClim**
BIO16	**Precipitation of Wettest Quarter (mm)**	**WorldClim**
BIO17	Precipitation of Driest Quarter	WorldClim
BIO18	**Precipitation of Warmest Quarter (mm)**	**WorldClim**
BIO19	Precipitation of Coldest Quarter	WorldClim
Alfisols	**Percent contribution**	**STATSGO2**
Andisols	Percent contribution	STATSGO2
Aridisols	Percent contribution	STATSGO2
Entisols	**Percent contribution**	**STATSGO2**
Gelisols	Percent contribution	STATSGO2
Histosols	**Percent contribution**	**STATSGO2**
Inceptisols	**Percent contribution**	**STATSGO2**
Mollisols	**Percent contribution**	**STATSGO2**
Oxisols	Percent contribution	STATSGO2
Spodosols	**Percent contribution**	**STATSGO2**
Ultisols	**Percent contribution**	**STATSGO2**
Vertisols	Percent contribution	STATSGO2

2.4. Ecological niche comparison analysis

To examine if the distributions of the associated environmental factors differ between our three subsets of data (GBIF Large, GBIF Small, and CMC/VSC), the distribution values of the environmental factors were extracted from the occurrence locations for each of the three independent datasets. These values were tested using the non-parametric two-sample Kolmogorov-Smirnov test (Lilliefors, 1967; Massey, 1951) in R using the function "ks.boot" with 1000 repetitions (Abadie, 2002; Sekhon, 2011). This non-parametric test does not require consistent sample size or assumptions of homogeneity of variance in the data.

2.5. Species distribution modelling

The program Maxent 3.3.3k (Phillips et al., 2006) was used to model potential suitable habitat of each of five species. Maxent uses species occurrence records and environmental factors to build a potential distribution across a defined geographic space, where each grid cell contains an index of habitat suitability (Newbold et al., 2009; Phillips et al., 2006). This model is specifically designed to work with presence-only datasets common in historical collections data and has been shown to consistently perform well when compared to other models (Elith et al., 2006; Phillips et al., 2004). Individual models were run for each of five species that consisted of the suite of herbaria data ('GBIF Large'), GBIF Large and Small herbaria data ('GBIF All', consisting of both subsets GBIF Large and GBIF Small), and GBIF Large and Small herbaria data plus CMC/VSC herbaria data ('GBIF CMC/VSC', consisting of all three

subsets).

To address potential spatial or environmental bias associated with NHC specimen-based data (Phillips et al., 2009), a bias file per dataset was created using the presence data of all Fuireneae species based on the assumption that all Fuireneae species were being searched during a collecting event of any given species. All occurrence records present within a grid cell of the same resolution and extent as the environmental layers were summed in ArcMap. Cells that had no occurrence records in them (NoData) were assigned the value of 0.1 (Kramer-Schadt et al., 2013) to indicate minimal sampling effort before exporting the final three bias files.

The fourteen selected climatic and soil variables were used as environmental layers (Table 1). Linear, quadratic, product, and hinge features were selected for all continuous variables; threshold functions were removed as hinging threshold functions produced unrealistic species-environment relationships (Heumann, 2013). Default values for remaining parameters were used following Newbold et al. (2009) with a regularization value of 1, convergence threshold of 0.00001, and a sample of 10,000 points to characterize the background points. All species data sets underwent 100 replicates of 1000 maximum iterations. Each replicate used a subset of 75% of records to calibrate the model, with the remainder of the records (25%) randomly subsampled with replacement for each replicate to validate the model. The area under the receiver operating characteristic curve (AUC), a threshold-independent metric, was used to examine model output compared to random background conditions (Phillips et al., 2004).

2.6. Difference in geographical predictions

To evaluate the niche overlap between models, ENMTools (Warren et al., 2008) was used to calculate Schoener's *D* and *I* indices for comparisons of each of the species models. Both indices analyze the similarity present between predictions of the three dataset inputs. They provide a direct comparison between the models, with values ranging from 0 (no overlap between niche models) to 1 (identical niche models; Warren et al., 2008).

Geographic differences in predicted distributions between the species models were calculated in ArcMap by the subtraction of output maps: GBIF CMC/VSC minus GBIF ALL (= contribution of CMC/VSC collections data), GBIF CMC/VSC minus GBIF Large (= contribution of small collections data overall), and GBIF ALL minus GBIF Large (= contribution of small collections data available in GBIF). Percent change values were classified into five classes showing how the habitat suitability index changed in the minuend as input datasets increased with small collections based data: 10–90% increase, 2–10% increase, 0–2% overall change, 2–10% decrease, and 10–90% decrease. The Albers equal-area conic projection was used for all resulting maps.

2.7. Change in habitat suitability

To assess the relative impact of the three species occurrence datasets on habitat suitability, resulting maps from Maxent were compared to the subsets of species occurrence data. Points in the three species occurrence datasets of the GBIF Large, GBIF Small (if applicable), and CMC/VSC collections data were used to extract the habitat suitability index in ArcMap at each point location from the up to three Maxent suitability maps produced using the data partitions for each species. Extracted indices were compared using the Mann-Whitney *U* test in R within a species dataset for species that only had records from GBIF Large and CMC/VSC collections, and the Kruskal-Wallis test for species that had records from each of the three datasets.

Table 2
Number of species occurrences for the three separate datasets of large collections data from GBIF (GBIF Large), small collections data from GBIF (GBIF Small), and data from the CMC and VSC herbaria (CMC/VSC) used as occurrence records for the Maxent modelling of Fuireneae species geographic niches. Datasets had duplicate records removed that occurred in the same grid cell as specified by the resolution of the environmental variables in ENMTools. Cells under "GBIF Small" for the first two species containing "n/a" did not have species occurrences that met qualifications within small herbaria from GBIF.

Species	GBIF large	GBIF small	CMC/VSC	Total
Fuirena squarrosa	25	n/a	44	69
Schoenoplectiella purshiana	45	n/a	15	60
Schoenoplectus acutus	434	52	13	499
Schoenoplectus pungens	413	32	26	471
Schoenoplectus tabernaemontani	352	38	29	419
Total	1269	122	127	1518

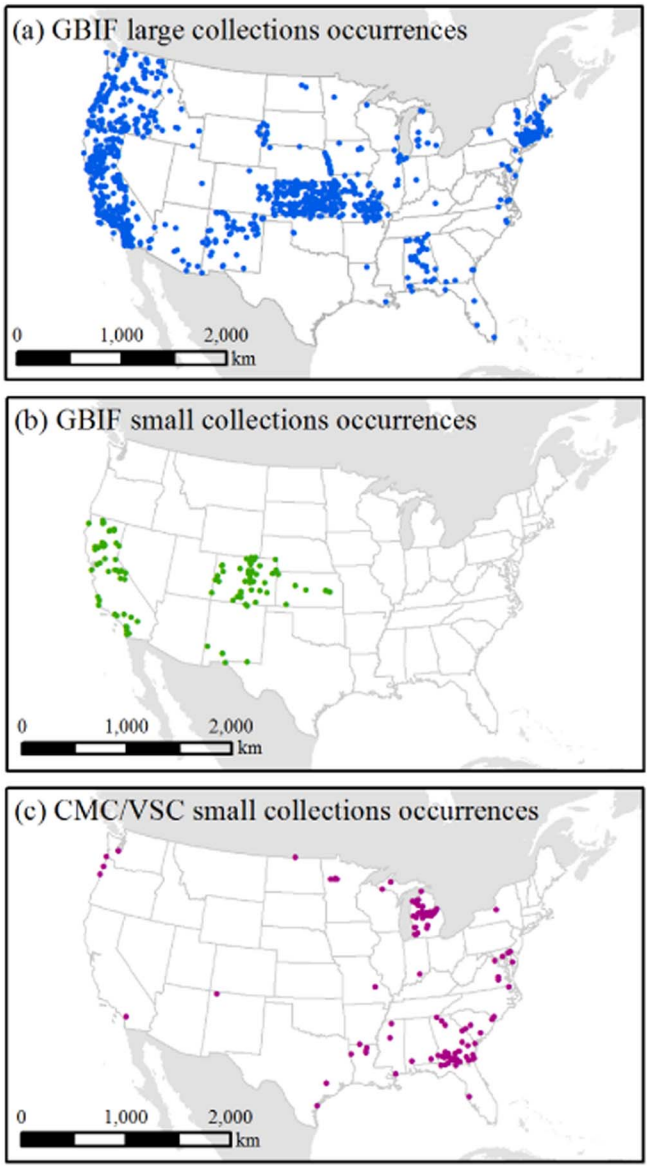

Fig. 2. Locations of occurrence records used for modelling from a) large collections from GBIF (n = 1269), b) small collections from GBIF (n = 122), and c) CMC and VSC small herbaria (n = 127). Each colored point represents one occurrence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Species occurrence datasets with extracted means and standard deviations of the environmental factor values. Includes the Kolmogorov-Smirnov *p*-value significance level calculated using ks.boot in R to compare the ecological niches of the datasets. All *p*-values are designated as > 0.05 (ns, for no significance), < 0.05, < 0.01, and < 0.001.

Species	Datasets compared by the Kolmogorov-Smirnov Test	Mean values and Kolmogorov-Smirnov *p*-value level of statistical significance for each comparison by environmental factor													
		Temp. seasonality (°C, Std dev ∗ 100)	Precip. of wettest quarter (mm)	Precip. of warmest quarter (mm)	Precip. seasonality (mm, coefficient of variation)	Mean temp of wettest quarter (°C, ∗10)	Mean diurnal range (°C, ∗10)	Max temp. of warmest month (°C, ∗10)	Alfisols	Entisols	Histosols	Inceptisols	Mollisols	Spodosols	Ultisols
Fuirena squarrosa	GBIF large	7220.6 ± 973.47	404.76 ± 59.99	361.08 ± 66.38	19.72 ± 4.-17	178.8 ± 7-8.34	123.2 ± 1-3.67	319.08 ± 20.04	9.08 ± 19.-44	14.48 ± 2-4.89	1.4 ± 2.95	8.16 ± 14-.41	1 ± 3.93	0.56 ± 2.8	59.84 ± 37.88
	CMC/VSC	6477.02 ± 496	414.09 ± 43.67	393 ± 66.-99	24.59 ± 6.-18	224.97 ± 62.84	128.72 ± 8.98	330.77 ± 5.37	6.45 ± 16.-08	11.77 ± 2-2.03	2.11 ± 10.-67	2.63 ± 5.-57	0 ± 0	2.75 ± 10-.13	71.59 ± 35.14
	p	< 0.001	ns	< 0.01	< 0.001	< 0.01	ns	< 0.001	ns	ns	ns	ns	ns	ns	ns
Schoenoplectiella purshiana	GBIF large	8690.97 ± 477.03	331.02 ± 35.6	288.93 ± 17.75	9.97 ± 4.6-7	77.17 ± 5-9.02	115.08 ± 9.19	281.75 ± 14.1	5.24 ± 18.-98	14.06 ± 2-1.26	4.2 ± 6.52	58.04 ± 3-8.01	0.2 ± 0.84	2.11 ± 12-.83	11.44 ± 29.8
	CMC/VSC	8786.06 ± 1118.33	328.33 ± 73.01	291.53 ± 37.06	20 ± 6.97	159.73 ± 58.95	119.86 ± 9	291.13 ± 20.72	29.53 ± 3-5.33	7.2 ± 20.-12	3.26 ± 5.7-7	2.86 ± 4.-64	11.4 ± 26.-28	6.6 ± 20.-36	35.13 ± 44.97
	p	< 0.05	< 0.05	ns	< 0.001	< 0.001	ns	ns	< 0.001	ns	< 0.05	< 0.001	< 0.01	ns	< 0.05
Schoenoplectus acutus	GBIF large	6577.58 ± 2134.36	295.78 ± 177.29	90.01 ± 1-01.44	66.12 ± 2-0.74	101.31 ± 76.38	140.99 ± 23.05	301.7 ± 4-1.59	16.26 ± 2-6.06	18.66 ± 2-6.46	0.82 ± 6.5-7	11.88 ± 2-2.65	28.39 ± 3-3.66	1.1 ± 7.34	0.99 ± 7-.54
	GBIF small	6588.96 ± 1569.2	257.69 ± 109.24	76.34 ± 7-2.84	66.13 ± 1-7.34	95.11 ± 7-9.3	150.84 ± 19	296.28 ± 41.63	11.8 ± 22.-47	20.25 ± 2-5.53	1.15 ± 8.3-2	9.17 ± 20-.39	33.67 ± 3-3.86	0 ± 0	0 ± 0
	CMC/VSC	9786.76 ± 1824.02	289.84 ± 73.84	241.07 ± 54.04	34.3 ± 15.-83	162 ± 40.-49	115.46 ± 8.51	268.76 ± 10.59	23.53 ± 3-1.3	25.07 ± 3-5.44	10.61 ± 2-2.52	8.38 ± 16-.05	13.76 ± 2-9.17	17 ± 31.3-2	0.15 ± 0-.55
	p GBIF large vs. CMC/VSC	< 0.001	< 0.05	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	ns	ns	< 0.001	ns	< 0.05	< 0.001	ns
	p GBIF large vs. GBIF small	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
	p GBIF small vs. CMC/VSC	< 0.001	< 0.01	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	ns	ns	< 0.001	< 0.05	< 0.05	< 0.001	ns
Schoenoplectus pungens	GBIF large	7685.42 ± 2241.38	265.34 ± 161.62	152.96 ± 111.13	53.77 ± 2-3.98	138.01 ± 81.85	142.03 ± 25.09	310.04 ± 42.38	14.42 ± 2-6.66	21.01 ± 2-7.58	0.39 ± 3.6-9	10.19 ± 2-1.84	31.37 ± 3-6.81	0.47 ± 3.-89	5 ± 18.7-3
	GBIF small	7003.18 ± 2145.55	206.18 ± 116.27	107.21 ± 63.23	61.75 ± 1-7.43	143.46 ± 68.32	153.09 ± 22.41	289.78 ± 40.47	12.12 ± 2-0.18	21 ± 26.3-6	0 ± 0	4.59 ± 9.-38	37.25 ± 3-4.22	0 ± 0	2 ± 10.0-9
	CMC/VSC	8752.8 ± 2098.03	340.96 ± 219.91	253.34 ± 66.43	28.73 ± 1-2.78	170.19 ± 49.86	112.57 ± 11.12	277.46 ± 32.8	33.11 ± 3-4.61	7.96 ± 13-.53	2.61 ± 4.6-3	12.3 ± 23-.93	8 ± 23.22	18.26 ± 2-7.42	10.65 ± 29.33
	p GBIF large vs. CMC/VSC	< 0.001	< 0.001	< 0.001	< 0.001	< 0.01	< 0.001	< 0.001	< 0.01	< 0.05	< 0.001	ns	< 0.001	< 0.001	ns
	p GBIF large vs. GBIF small	< 0.01	< 0.001	< 0.001	< 0.05	< 0.05	< 0.002	< 0.001	ns	ns	ns	ns	ns	ns	ns
	p GBIF small vs. CMC/VSC	< 0.001	< 0.001	< 0.001	< 0.001	ns	< 0.003	< 0.01	< 0.05	< 0.05	< 0.01	ns	< 0.001	< 0.001	ns
Schoenoplectus tabernae-montani	GBIF large	8530.17 ± 1707.92	330.81 ± 133.32	241.24 ± 107.71	40.6 ± 23.-22	145.38 ± 81.78	131.95 ± 20.18	304.63 ± 28.93	11.67 ± 2-5.24	12.33 ± 2-2.24	0.96 ± 3.5-3	20.14 ± 3-3.83	38.45 ± 4-1.95	1.55 ± 9.-72	6.71 ± 2-1.94
	GBIF small	8573.65 ± 759.8	184.39 ± 59.59	154.05 ± 57.27	50.92 ± 1-5.12	164.76 ± 44.53	165.21 ± 14.28	301.55 ± 47.8	7.36 ± 20.-05	24.76 ± 2-5.63	0 ± 0	5.52 ± 11-.26	35.71 ± 2-3.61	0 ± 0	0 ± 0
	CMC/VSC	8845.03 ± 2190.69	310.03 ± 159.74	251.75 ± 101.87	33.17 ± 1-8.61	182.51 ± 49.95	117.79 ± 13.7	283.37 ± 28.1	33.17 ± 3-5.52	15.34 ± 2-6.91	4.58 ± 14.-01	9.37 ± 19-.05	11.48 ± 2-3.63	8.62 ± 21-.98	15.03 ± 32.54
	p GBIF large vs. CMC/VSC	ns	< 0.001	ns	< 0.01	< 0.01	< 0.001	< 0.001	< 0.01	ns	< 0.01	ns	< 0.01	< 0.001	< 0.05
	p GBIF large vs. GBIF small	< 0.05	< 0.001	< 0.001	< 0.01	< 0.05	< 0.001	ns	ns	< 0.01	< 0.01	< 0.05	ns	ns	ns
	p GBIF small vs. CMC/VSC	< 0.001	< 0.001	< 0.001	< 0.001	< 0.05	< 0.001	< 0.01	< 0.001	ns	< 0.001	ns	< 0.001	< 0.001	< 0.01

3. Results

3.1. Number of records used in modelling

A total of five species were included in the analyses (Table 2). The total number of Fuireneae records originally returned from the GBIF database was 15,967, of which 13,411 (84%) originated from vouchered herbarium specimens. Of vouchered specimens, a total of 5164 (34%) records included a georeferenced locality. The final reduced GBIF data set consisted of 1391 records from the three subsets: 1269 from large collections (Fig. 2a) and 122 from small collections (Fig. 2b). Records from the CMC/VSC collections totaled 127 (Fig. 2c, Table 2).

3.2. Ecological niche comparison analysis

The majority of all comparisons among datasets for each of the five species had a minimum of five significantly different environmental factors, with 62.34% of all dataset comparisons showing significantly different environmental values between datasets (Table 3, Appendix B, Figs. B.1–B.5). Of the significantly different comparisons, 41.67% were soil factors and 58.33% were climatic variables. One species, *S. acutus*, did not have significant differences in soil or climatic variable between the records from GBIF Large and GBIF Small datasets, but did have 10 and 11 significant soil and climatic differences when the CMC/VSC collections datasets were compared to GBIF Large and GBIF Small datasets, respectively. *Schoenoplectus tabernaemontani* had the highest number of significant soil and climatic differences (73.81%) among the three-dataset comparisons.

3.3. Model results

Output maps display habitat suitability indices for the models based on data from GBIF Large, GBIF All, and GBIF CMC/VSC collections data (see Figs. 3-7). The habitat suitability indices for *F. squarrosa* lessened throughout the western half of the United States when CMC/VSC collections data were added, as is congruent with the currently known distribution, though the majority of cells experienced little to no change in the habitat suitability index. The habitat suitability indices for *S. purshiana* showed greater change in the mid-eastern U.S. regions. *Schoenoplectus acutus*, *S. pungens*, and *S. tabernaemontani* are widely distributed species across the United States, and all had distinct variations present across the background suitability indices. The mean testing AUC values from the Maxent model for each species dataset ranged from 0.655 to 0.933 (Appendix A, Table A.3).

3.4. Difference in geographical predictions

All models became less similar (both D and I) as more small collections data were added to the input (Table 4). The greatest differences were seen between the models when examining data for *F. squarrosa* ($D = 0.826$, $I = 0.971$) and *S. purshiana* ($D = 0.745$, $I = 0.909$); both species had no small collections data from GBIF added, occupy a narrower range in the United States than the *Schoenoplectus* species, and were comparatively heavily sampled in the CMC/VSC collections.

The comparisons between the two models of GBIF & CMC/VSC collections data and GBIF Large collections data showed the greatest overall differences in every species by percent (Figs. 3-7, Supplementary material Appendix A, Table A.4). For species that had three datasets to compare, the comparison between the models based on GBIF & CMC/VSC collections data and based on GBIF Large collections data had the lowest number of cells in the class of 0–2% change in cell predictions (21.54% - 28.70%). This was lower than the two remaining model comparisons (GBIF Small collections data vs. GBIF Large collections data and GBIF Small collections data vs. GBIF & CMC/VSC collections data) that had a calculated 0–2% change in cell predictions of 28.18% - 34.87%. The two species, *F. squarrosa* and *S. purshiana*, with the single comparison of GBIF & CMC/VSC data and GBIF Large data had a smaller difference between models as 55.94% to 64.19% of cells

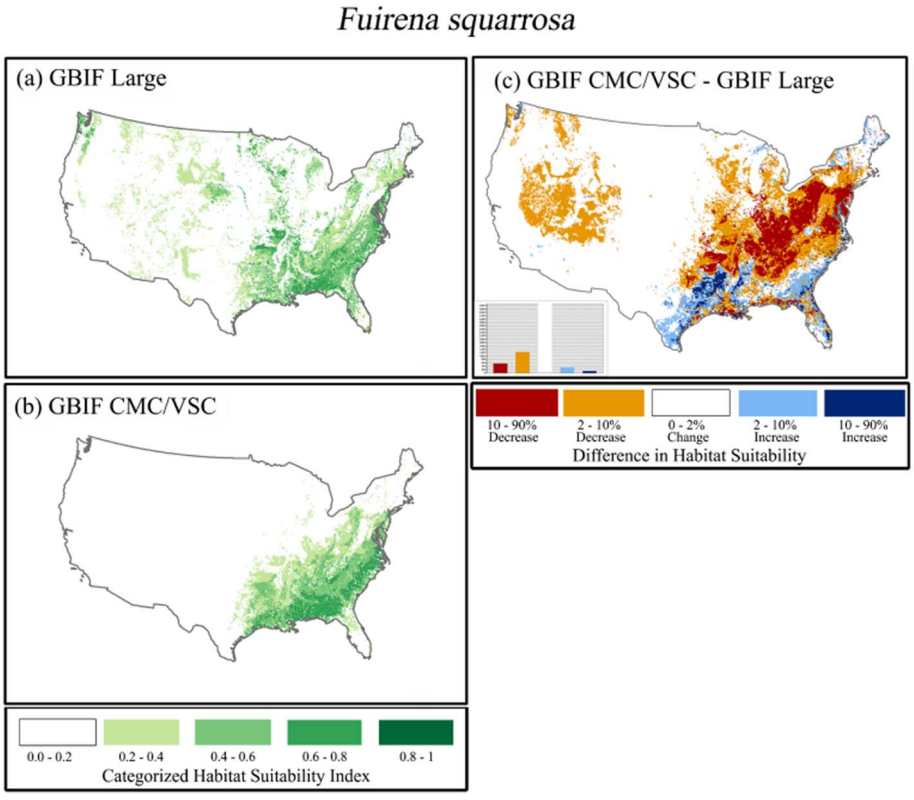

Fig. 3. Geographic differences between the maps of the habitat suitability index for the following model comparisons: GBIF & CMC/VSC data – GBIF Large data for *Fuirena squarrosa*. Colors in maps and legends correspond to the categorized percent increase or decrease in grid cells that the first model output listed experienced after the subtraction of the second model. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

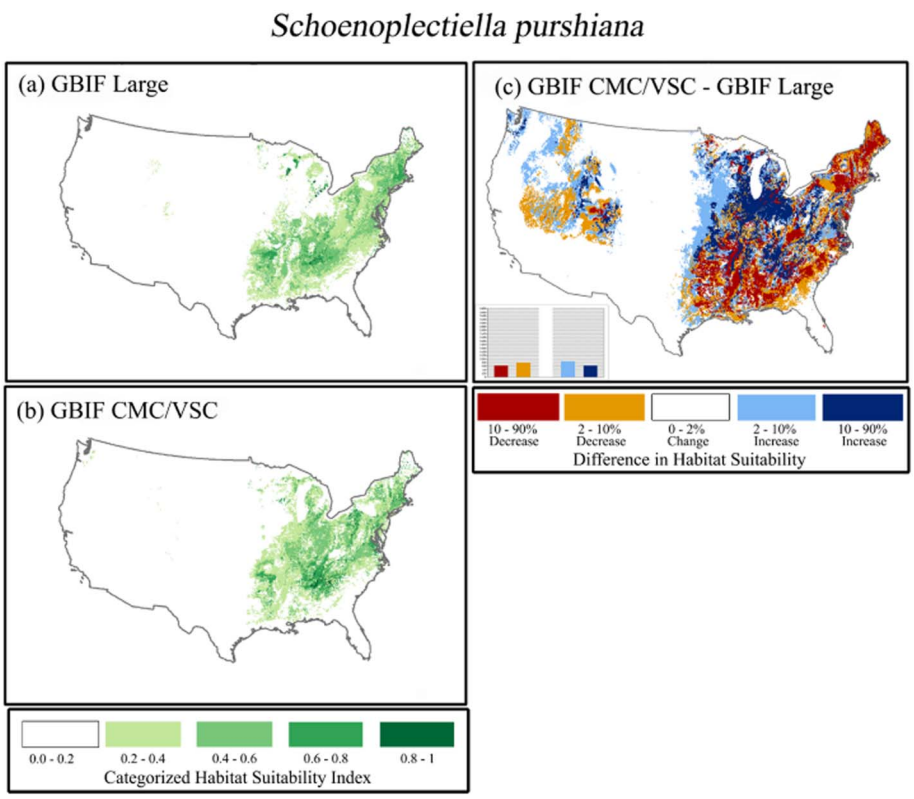

Fig. 4. Geographic differences between the maps of the habitat suitability index for the following model comparisons: GBIF & CMC/VSC data – GBIF Large data for *Schoenoplectiella purshiana*. Colors in maps and legends correspond to the categorized percent increase or decrease in grid cells that the first model output listed experienced after the subtraction of the second model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

had a 0–2% change; however, large differences still occurred between models.

Schoenoplectus tabernaemontani had the largest number of cells with differences > 2% in predictions. The comparison between the model based on the GBIF Large collections dataset and the model based on GBIF & CMC/VSC collections datasets had the highest percentage of cells (19.33% increase and 3.97% decrease) displaying over a 10% change, whereas the remaining two model comparisons had fewer cells (9.4% increase and 4% decrease) displaying over a 10% increase or decrease. *Fuirena squarrosa* had the lowest percent change overall with 64.19% of cells experiencing less than a 2% difference.

3.5. Change in habitat suitability

The small collections data when added to the models did not increase the habitat suitability indices for the applied GBIF Large dataset (Appendix A, Table A.5). However, the mean habitat suitability indices for the applied CMC/VSC dataset increased when GBIF Small collections were included in the input. The mean habitat suitability index for the applied CMC/VSC occurrences increased by 0.07–0.25 for all species with the input of CMC/VSC data.

4. Discussion

4.1. Influence of small collections data on habitat suitability of Fuireneae

The regional representation present in small collections data is an important source of data with potential to enhance the value of biodiversity studies. Both the characterized environmental ranges and the resulting models of suitable habitat were markedly affected by the addition of data from small collections, showing the impact that even two small herbaria (CMC & VSC) can have on SDMs. Increasing the number of species occurrences through specific inclusion of data from small collections produced a model built on a more robust definition of the species environmental ranges.

Numerous studies have suggested that adding a higher number of input occurrences increases model performance (see Cumming, 2000; Hernandez et al., 2006; Stockwell and Peterson, 2002; Wisz et al., 2008); however, by partitioning herbarium data into classes based on size attributes of their sources, our study indicates the nature of the data source can also be an important factor. Including data from small collections alongside data from large collections capitalizes on the robust regional sampling of diversity typical of small collections and together with data from large collections the combined datasets are able to produce an adjusted, greater understanding of species distributions. Three of the species (*S. acutus*, *S. pungens*, and *S. tabernaemontani*) used in the study had > 350 records in the GBIF Large collection dataset, with fewer than 40 records added cumulatively from both GBIF Small and the CMC/VSC small collection datasets. All three of these species experienced significant differences among models based on datasets of large and small collections despite the relatively small percentage of occurrence points contributed from small collections to the occurrence dataset. If a similar range and environmental conditions are sampled by multiple datasets, then the impact of combining these datasets will be less apparent than when combining multiple datasets which sample different areas of the species range. In this study, data from small collection datasets (GBIF Small and CMC/VSC) complimented the GBIF Large data set, and as they represent occurrence records with unique environmental data, data from small collections is essential for the development of a robust and comprehensive species distribution model.

The AUC indicates non-randomness of the habitat suitability output, and is an assessment metric that was used to compare the model

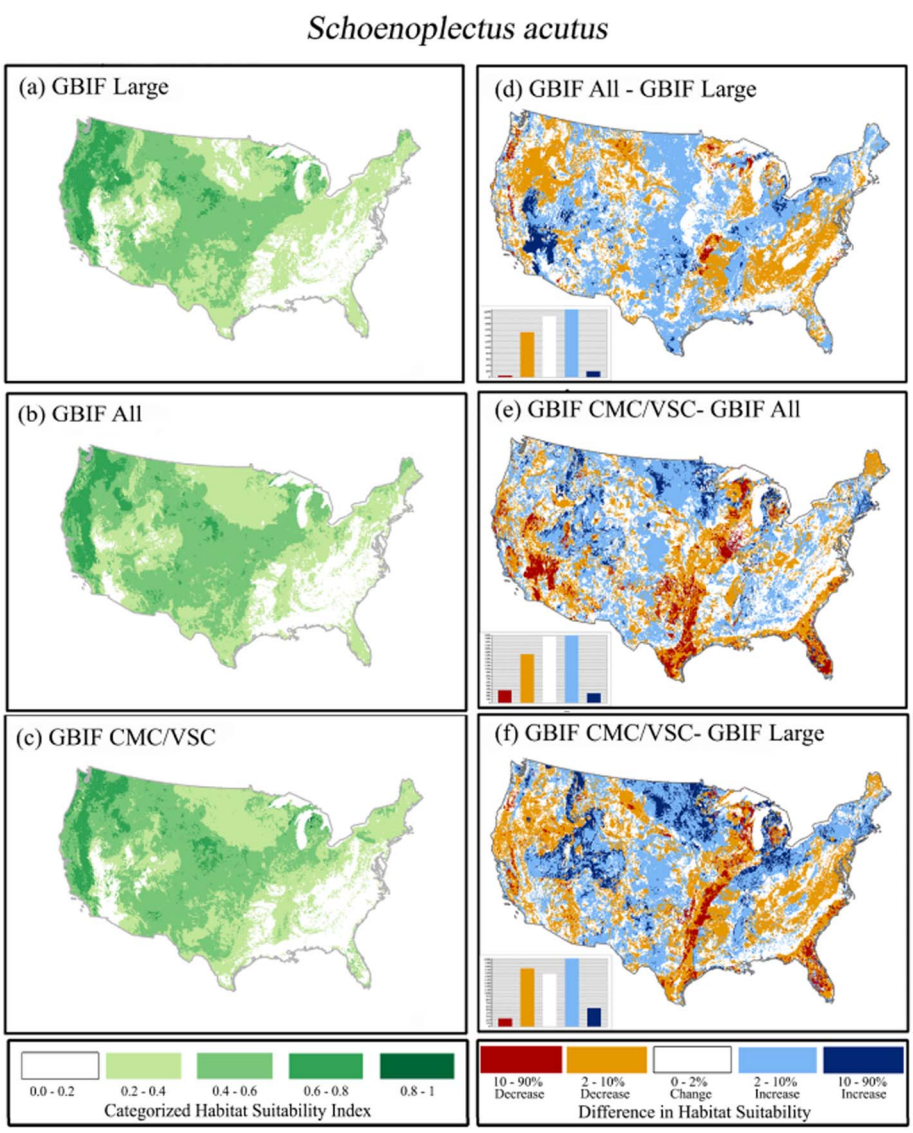

Fig. 5. Geographic differences between the maps of the habitat suitability index for the following model comparisons: GBIF All data – GBIF Large data, GBIF CMC/VSC data – GBIF All data, and GBIF & CMC/VSC data – GBIF Large data for *Schoenoplectus acutus*. Colors in maps and legends correspond to the categorized percent increase or decrease in grid cells that the first model output listed experienced after the subtraction of the second model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

outputs with a random model. Generalist species will tend to have a lower AUC than specialist species based on their inherent distributions. The AUC (Appendix A, Table A.3) for three species, *S. purshiana, S. pungens*, and *S. tabernaemontani*, decreases consistently as more data from small collections are added. Since generalist species such as *S. pungens* and *S. tabernaemontani* are able to survive in a wider niche, the addition of more occurrence points results in the expansion of the modelled niche (thus encompassing a greater range of environmental characteristics). Predictably, the result is more random model predictions (Lobo et al., 2008). Conversely, *F. squarrosa* inhabits a smaller niche, and as more occurrences are added to the model inputs, the AUC rises indicating the modelled habitat becomes more specific.

In further support of small collections providing a valuable data resource for species distribution modelling, the addition of small collections to model inputs in this study resulted in the ability of models to better predict other small collections data (Appendix A, Table A.5, see *S. pungens* and *S. tabernaemontani*). While this seems obvious and may lack statistical rigor, it supports the assertion that omitting data from small collection datasets can negatively affect model results. In terms of Schoener's *D* and *I* statistics, none of the five species had an identically modelled niche, and the addition of more small collections decreased the overlap. The species with a narrow range in the United States, *F. squarrosa* and *S. purshiana*, had the least overlap while the remaining widespread species in *Schoenoplectus* had above 0.9 for both the *D* and the *I*. Although the *Schoenoplectus* species had a greater number of occurrences, as seen with the environmental value assessment, increased sample size alone does not produce a more accurate or precise model. The three species that contributed over 350 occurrence points from the GBIF Large collections (*S. acutus*, *S. pungens*, and *S. tabernaemontani*) were the most widely distributed species in this study, making them inherently difficult for modelling as they have a wide range of habitats that are challenging to precisely define as a much larger sample size is required (Hernandez et al., 2006).

The small collections datasets used within this study contributed a disproportionately large amount of unique information for the building of our model. These data are able to fill in gaps in our understanding of the ecological parameters under which these species live, thus refining the model and mapping with greater accuracy. With narrowly endemic

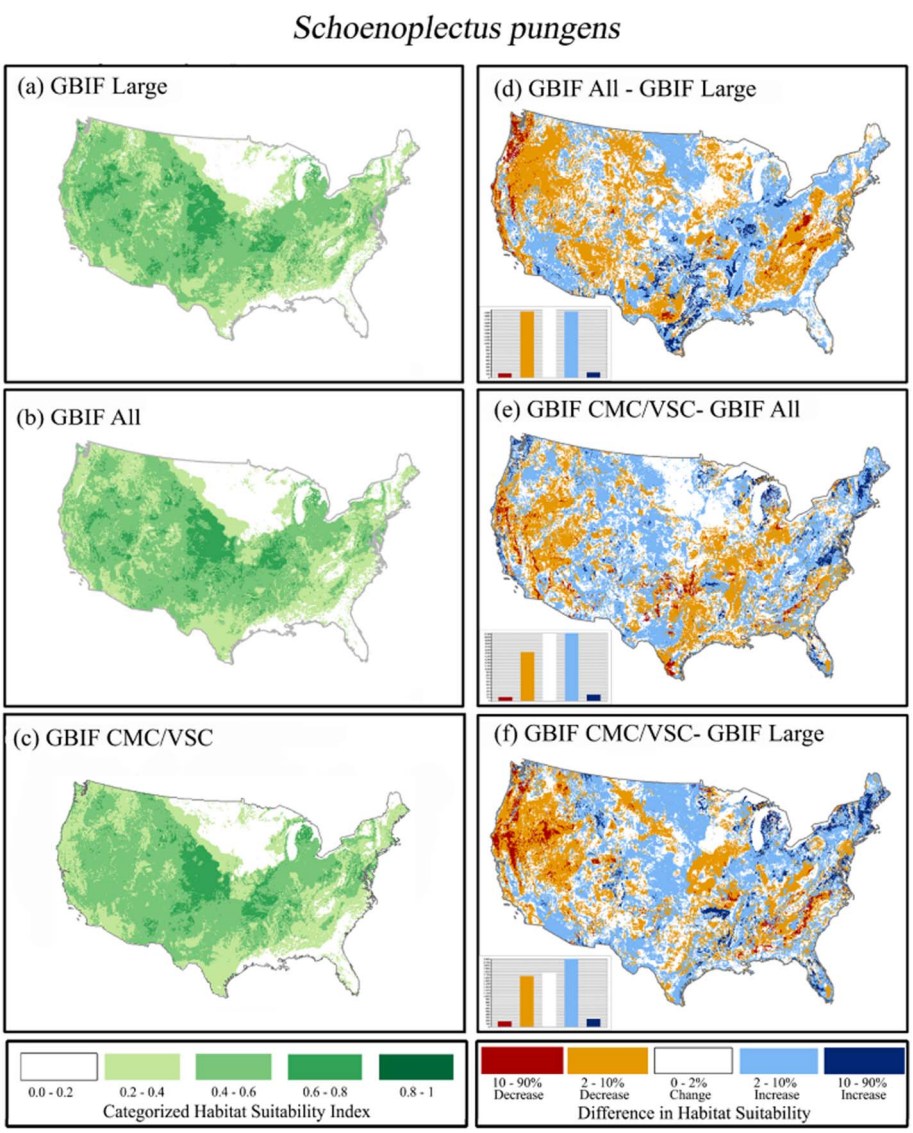

Fig. 6. Geographic differences between the maps of the habitat suitability index for the following model comparisons: GBIF All data – GBIF Large data, GBIF CMC/VSC data – GBIF All data, and GBIF & CMC/VSC data – GBIF Large data for *Schoenoplectus pungens*. Colors in maps and legends correspond to the categorized percent increase or decrease in grid cells that the first model output listed experienced after the subtraction of the second model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

species, such as *S. purshiana*, a smaller sample size could suffice if it is able to encompass on a regional scale all ecological characteristics of that species for the input environmental factors. The increase of 0.25 in the mean habitat suitability index for the *S. purshiana* CMC/VSC dataset, combined with the lowest niche overlap of the five species, demonstrated that the addition of data from small collections, both from GBIF small NHCs and the CMC/VSC datasets, captured unique geographical and ecological data for the species that was not present in the large NHC datasets available through GBIF.

This study is intended as a case study to demonstrate the value of data available in small collections when building species distribution models from NHC's. As with all species distribution models, limitations persist in the resulting models. We focused this analysis on the potential impact and contribution of small collections data to niche-based modelling using the widely used and accepted Maxent model as an exemplar. Our analyses are based on reliable and easily available climate and soils data at a relatively coarse grain to describe and define species niches. Though finer grain data and the additional of other environmental factors such as land use histories and spatial landscape relationships also affect species distributions, they were not included in these analyses due to increased model complexity and lack of comprehensive available data.

Small collections are rich resources for unduplicated specimens representing intense regional, temporal, and community sampling, creating a “hidden source” of specimens (Nelson and Monfils, 2015). This is reflected in the contributed data from CMC and VSC collections. The CMC and VSC collections represent a high diversity of Fuireneae taxa. The regional CMC and VSC physical collections are in close proximity to areas of high diversity for the tribe and their respective herbarium directors have research focused on Fuireneae species. This results in directed collections of the regional narrow endemic species within the tribe. This typifies a trend found for the taxonomic diversity of small collections. Though each collection may not individually contribute substantially to all models across all taxa, once all small collections are mobilized in accessible databases, they will together have vast potential to improve species distribution models through their fine scale and directed sampling.

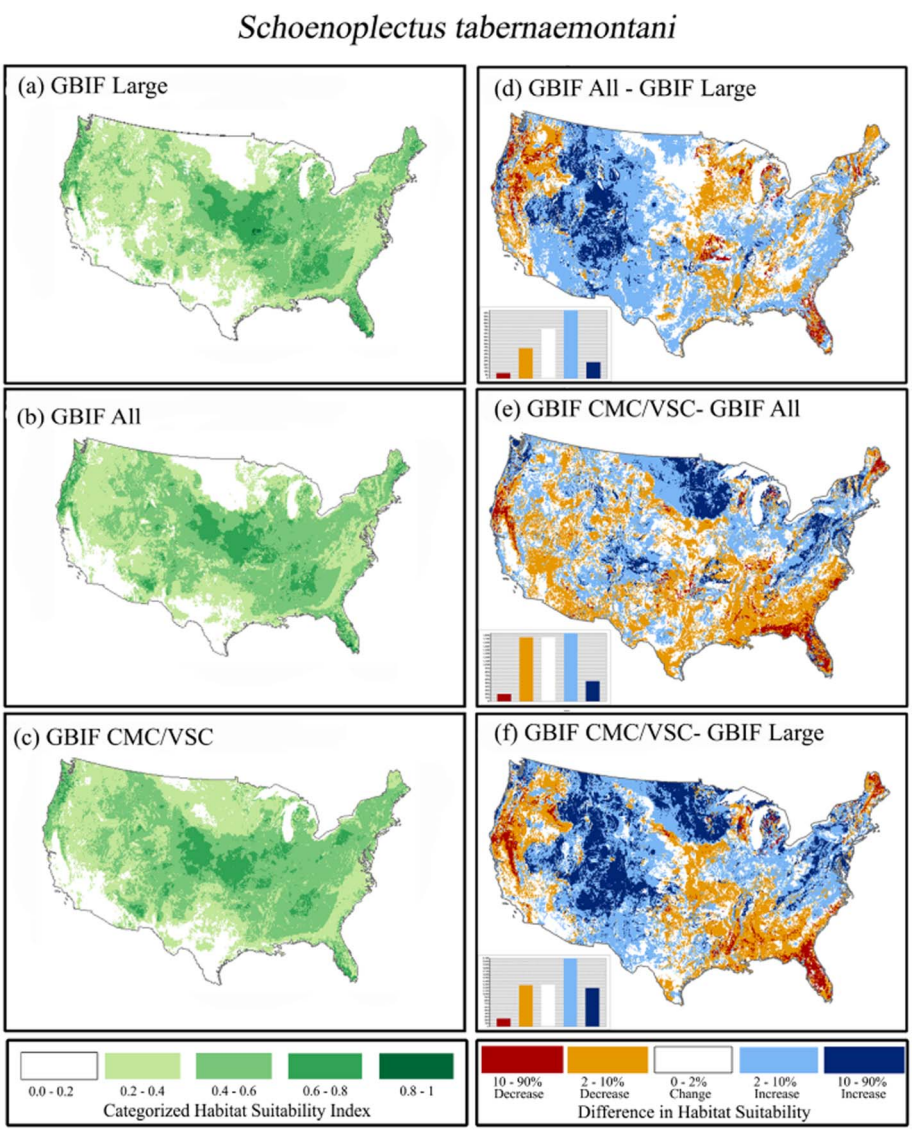

Fig. 7. Geographic differences between the maps of the habitat suitability index for the following model comparisons: GBIF All data – GBIF Large data, GBIF CMC/VSC data – GBIF All data, and GBIF & CMC/VSC data – GBIF Large data for *Schoenoplectus tabernaemontani*. Colors in maps and legends correspond to the categorized percent increase or decrease in grid cells that the first model output listed experienced after the subtraction of the second model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Conclusions

Online data portals, such as the expanding GBIF portal (based on both collections and observation based data) and iDigBio (based solely on collections data), are critical as they increase the accessibility of freely available biodiversity data from herbaria and natural history collections, enabling efficient extraction and compilation of data from numerous sources into a standard dataset. Currently, large natural history collections are the primary contributors in the digitization movement of natural history collections (Ariño, 2010; Gaiji et al., 2013). Increasing the availability of data from small collections in online databases will expand our understanding of biota, not only globally but on a highly refined localized scale.

This research has demonstrated the potential impact of small collections not only in terms of the geographic and environmental distribution added by small collections datasets, but also the impact these data can have when used in a species distribution models. We also demonstrated the importance of carefully working with these data to meet the stringent requirements for data analysis, particularly the geospatial attributes of the data. To increase the utility of biodiversity collections data, as digitization and specimen collection efforts continue, adherence to best georeferencing practices is essential particularly in small collections that are of manageable size (see Guralnick et al., 2007; Maldonado et al., 2015).

As a community, it is necessary for natural history collections curators to maintain networks that incorporate all types and sizes of collections. The contributions from all collections provide the resolution, expertise, and reliability to make more precise predictions of species distributions. This case study using Fuireneae aids in illustrating the potential value of small collections in creating reliable species distribution models and reinforces the need for inclusion of all collections data into publicly accessible databases such as GBIF and iDigBio. Inclusion of data from small regional collections has untapped potential to increase the resolution of scientific studies and enhance our understanding of global biodiversity.

Table 4
Niche overlap of models using Schoener's *D* and *I* statistic. A value of 0 indicates no niche overlap between models and 1 indicates identical modelled niches.

		Large GBIF		All GBIF		GBIF CMC/VSC	
		D	*I*	*D*	*I*	*D*	*I*
Fuirena squarrosa	Large GBIF	1	1	n/a	n/a	–	–
	GBIF CMC/VSC	0.826	0.971	n/a	n/a	1	1
Schoenoplectiella purshiana	Large GBIF	1	1	n/a	n/a	–	–
	GBIF CMC/VSC	0.745	0.909	n/a	n/a	1	1
Schoenoplectus acutus	Large GBIF	1	1	–	–	–	–
	All GBIF	0.95	0.997	1	1	–	–
	GBIF CMC/VSC	0.928	0.995	0.939	0.996	1	1
Schoenoplectus pungens	Large GBIF	1	1	–	–	–	–
	All GBIF	0.952	0.998	1	1	–	–
	GBIF CMC/VSC	0.944	0.997	0.952	0.998	1	1
Schoenoplectus tabernaemontani	Large GBIF	1	1	–	–	–	–
	All GBIF	0.935	0.996	1	1	–	–
	GBIF CMC/VSC	0.912	0.993	0.935	0.996	1	1

Acknowledgements

Thanks to SPNHC 2015 for awarding funds for travel to present this work at its annual conference, and to the CMU Biology department for graduate assistantship funds to work on this study. Thanks to Gil Nelson for his comments and feedback on our manuscript. Thanks to John Gross for assistance in editing soil data layers, to Daniel Spalink, Rob Guralnick, Kevin Pangle, Charlotte Germain-Aubrey, Mael Glon, Clint Pogue, and Rachel Hackett for valuable suggestions and support in the creation of the project. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2017.09.009.

References

Abadie, A., 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. J. Am. Stat. Assoc. 97, 284–292.
Ariño, A.H., 2010. Approaches to estimating the universe of natural history collections data. Biodivers. Inform. 7, 81–92.
Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200, 1–19.
Beach, J., et al., 2012. Implementation plan for the network integrated biocollections alliance. In: Workshop Report.
Beaman, R.S., Cellinese, N., 2012. Mass digitization of scientific collections: new opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys 17, 7–17.
Beck, J., et al., 2013. Online solutions and the "Wallacean shortfall": what does GBIF contribute to our knowledge of species' ranges? Divers. Distrib. 19, 1043–1050.
Beck, J., et al., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Eco. Inform. 19, 10–15.
Berendsohn, W.G., et al., 2010. Recommendations of the GBIF Task Group on the global strategy and action plan for the mobilization of natural history collections data. Biodivers. Inform. 7, 67–71.
Breiner, F.T., et al., 2015. Overcoming limitations of modelling rare species by using ensembles of small models. Methods Ecol. Evol. 6, 1210–1218.
Chapman, A.D., Speers, L., 2005. Uses of Primary Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility. Report, Copenhagen, Denmark.
Chauvel, B., et al., 2006. The historical spread of *Ambrosia artemisiifolia* L. in France from herbarium records. J. Biogeogr. 33, 665–673.
Crawford, P.H.C., Hoagland, B.W., 2009. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? J. Biogeogr. 36, 651–661.
Cumming, G.S., 2000. Using between-model comparisons to fine-tune linear models of species ranges. J. Biogeogr. 27, 441–455.
Davis, C.C., et al., 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. Am. J. Bot. 102, 1599–1609.
Dormann, C.F., et al., 2012. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.
Elith, J., et al., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.
Faith, D.P., et al., 2013. Bridging biodiversity data gaps: recommendations to meet users' data needs. Biodivers. Inform. 8, 41–58.
Ferro, M.L., Flick, A.J., 2015. "Collection bias" and the importance of natural history collections in species habitat modeling: a case study using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a critique of GBIF.org. Coleopt. Bull. 69, 415–425.
Flora of North America Editorial Committee, 2014. Flora of North America North of Mexico. New York and Oxford.
Gaiji, S., et al., 2013. Content assessment of the primary biodiversity data published through GBIF Network: status, challenges and potentials. Biodivers. Inform. 8, 94–172.
Gallagher, R.V., et al., 2009. Phenological trends among Australian alpine species: using herbarium records to identify climate-change indicators. Aust. J. Bot. 57, 1–9.
García-Roselló, E., et al., 2014. Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? Glob. Ecol. Biogeogr. 24, 335–347.
GBIF, 2015. 2014 GBIF Science Review. 52 pp. Global Biodiversity Information Facility, Copenhagen Available online at. http://www.gbif.org/2014-science-review.
Govaerts, R., et al., 2014. World Checklist of Cyperaceae Published Update. Trustees of the Royal Botanic Gardens, Kew Available online at. http://www.kew.org/wcsp/ [Retrieved 15 September 2015].
Graham, C.H., et al., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19, 497–503.
Guillera-Arroita, G., et al., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24, 276–292.
Guo, W.-Y., 2013. Invasion of Old World *Phragmites australis* in the New World: precipitation and temperature patterns combined with human influences redesign the invasive niche. Glob. Chang. Biol. 19, 3406–3422.
Guralnick, R.P., et al., 2007. Towards a collaborative, global infrastructure for biodiversity assessment. Ecol. Lett. 10, 663–672.
Heikkinen, R.K., et al., 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Prog. Phys. Geogr. 30, 751–777.
Hernandez, P.A., et al., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.
Heumann, B.W., 2013. Land suitability modeling using a geographic socio-environmental niche-based approach: a case study from Northeastern Thailand. Ann. Assoc. Am. Geogr. 103, 1199–1216.
Hijmans, R.J., et al., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.
Hortal, J., et al., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117, 847–858.
Kramer-Schadt, S., et al., 2013. The importance of correcting for sampling bias in Maxent species distribution models. Divers. Distrib. 19, 1366–1379.
Lavoie, C., 2013. Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. Perspect. Plant Ecol. Evol. Syst. 15, 68–76.
Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 62, 399–402.
Lobo, J.M., et al., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151.
Maldonado, C., et al., 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob. Ecol. Biogeogr. 24, 973–984.
Massey Jr., F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46, 68–78.
Meyer, C., Weigelt, P., Kreft, H., 2016. Multidemensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19, 992–1006.
Monfils, A.K., Nelson, G., 2014. The Small Collections Network: Recruiting, Retaining, and Sustaining Small Collections in Biodiversity Digitization. 28:1. Society for the Preservation of Natural History Collections: Academy Research, pp. 32–33.
Nelson, G., Monfils, A.K., 2015. SCNet: Supporting Digitization in Small Collections. Presentation: Society for the Preservation of Natural History Collections. pp. 2015.
Newbold, T., 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Prog. Phys. Geogr. 34, 3–22.
Newbold, T., et al., 2009. Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. J. Biogeogr. 36, 2085–2095.
Pearson, R.G., et al., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J. Biogeogr. 34, 102–117.
Phillips, S.J., et al., 2004. A maximum entropy approach to species distribution modeling. - proceedings of the twenty-first international conference on. Mach. Learn. 655–662.

Phillips, S.J., et al., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.
Phillips, S.J., et al., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.
Pyke, G.H., Ehrlich, P.R., 2010. Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biol. Rev. 85, 247–266.
Rios, N.E., Bart, H.L., 2010. GEOLocate (Version 3.22) [Computer Software]. Tulane University Museum of Natural History, Bell Chasse, LA.
Robbirt, K.M., et al., 2011. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid *Ophrys sphegodes*. J. Ecol. 99, 235–241.
Roberts, R.P., et al., 2015. Available at: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559.
Sekhon, J., 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J. Stat. Softw. 42.
Shiels, D.R., et al., 2014. Monophyly and phylogeny of *Schoenoplectus* and *Schoenoplectiella* (Cyperaceae): evidence from chloroplast and nuclear DNA sequences. Syst. Bot. 39, 142–144.
Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture Web Soil Survey (STATSGO2). Available online at. http://websoilsurvey.nrcs.usda.gov.
Stockwell, D.R.B., Peterson, T.A., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Model. 148, 1–13.
Suarez, A.V., Tsutsui, N.D., 2004. The value of museum collections for research and society. Bioscience 54, 66–74.
Theirs, B., 2014. Index Herbariorum: a global directory of public herbaria and associated staff. In: New York Botanical Garden's Virtual Herbarium, Available at: http://sweetgum.nybg.org/ih.
Warren, D.L., et al., 2008. Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. Evolution 62, 2868–2883.
Warren, D.L., et al., 2010. ENMTools: a toolbox for comparative studies of environmental niche models. Ecography 33, 607–611.
Wen, J., et al., 2015. Collections-based systematics: opportunities and outlook for 2050. J. Syst. Evol. 53, 477–488.
Wisz, M.S., et al., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773.
Yesson, C., et al., 2007. How global is the global biodiversity information facility? PLoS One 2, e1124.