Received: 15 November 2022 | Accepted: 28 November 2022

DOI: 10.1002/ecs2.4419

ARTICLE

ECOSPHERE
AN ESA OPEN ACCESS JOURNAL

Does adding community science observations to museum records improve distribution modeling of a rare endemic plant?

Andrew G. Gaier | **Julian Resasco**

Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA

Correspondence
Andrew G. Gaier
Email: andrewggaier@gmail.com

Funding information
Colorado Native Plants Society; National Geographic Society; National Science Foundation, Grant/Award Number: 2102974; University of Colorado Boulder

Handling Editor: Debra P. C. Peters

Abstract

Understanding the ranges of rare and endangered species is central to conserving biodiversity in the Anthropocene. Species distribution models (SDMs) have become a common and powerful tool for analyzing species–environment relationships across geographic space. Although evaluating the distribution of rare species is integral to their conservation, this can be difficult when limited distribution data are available. Community science platforms, such as iNaturalist, have emerged as alternative sources for species occurrence data. Although these observations are often thought to be of lower quality than those of natural history collections, they may have potential for improving SDMs for species with few occurrence records from collections. Here, we investigate the utility of iNaturalist data for developing SDMs for a rare high-elevation plant, *Telesonix jamesii*. Because methods for modeling rare species are limited in the literature, five different modeling techniques were considered, including profile methods, statistical models, and machine learning algorithms. The inclusion of iNaturalist data doubled the number of usable records for *T. jamesii*. We found that a random forest (RF) model using ensemble training data performed the highest of any model (area under curve = 0.98). We then compared the performance of RF models that use only natural history training data and those that use a combination of natural history (herbarium specimens) and iNaturalist training data. All models heavily relied on climate data (mean temperature of driest quarter, and precipitation of the warmest quarter), indicating that this species is under threat as climate continues to change. Validation datasets affected model fits as well. Models using only herbarium data performed slightly poorer when evaluated with cross-validation than when validated externally with iNaturalist data. This study can serve as a model for future SDM studies of species with similar data limitations.

KEYWORDS
alpine botany, biogeography, Chasmophyte, ecological niche model, iNaturalist, Rocky Mountains, Saxifragaceae, species distribution model

INTRODUCTION

Species distribution models (SDMs) have become a common tool for analyzing species–environment relationships and projecting range estimates across geographic space (Guisan et al., 2017). By using species occurrence data along with environmental predictors, these models can aid in species monitoring (Williams et al., 2009), designating conservation areas (Koch et al., 2017), and providing insights into how the distribution of a species may be affected by anthropogenic stressors such as biological invasion, human land use, and climate change (Franklin, 2013). Several studies suggest that these stressors have led to species range contractions and extinctions (Chen et al., 2011; Fois et al., 2018; Imperio et al., 2013; Parmesan & Yohe, 2003). Therefore, to mitigate declines of global biodiversity, research efforts directed toward modeling species at higher risks of extinction are critical (Lomba et al., 2010; Wang et al., 2012).

Rare and endemic plant species are particularly threatened and could benefit from applications of SDMs (Breiner et al., 2015; Guisan et al., 2013). Extinction risk is often higher for these species due to narrow geographic ranges, small population sizes, specialized habitat preferences, reduced gene flow, and higher vulnerability to environmental changes (Kruckeberg & Rabinowitz, 1985; Lavergne et al., 2005; Lomba et al., 2010; Mousikos et al., 2021). Therefore, there is great urgency for identifying and protecting habitat for rare species (Mousikos et al., 2021; Singh, 2013). However, despite being in the most need of predictive distribution modeling, rare species remain the most challenging to model. This challenge has been referred to as the "rare species modeling paradox" (Lomba et al., 2010).

Modeling the distributions of rare species is challenging for several reasons. SDMs rely on accurate datasets of georeferenced species occurrences (Fletcher & Fortin, 2018). Correlations between presence locations and environmental variables provide an estimate of a given species' fundamental niche across space (Guisan & Thuiller, 2005). Natural history collections have served as reliable sources for obtaining the distribution data necessary for these models. For rare species, these datasets are often limited, containing only a small number of occurrences gathered over long periods of time (Rushton et al., 2004). Small sample sizes may compromise model robustness if the full range of a species is not well represented in records (Pearson et al., 2007; Williams et al., 2009). Additionally, rare species are often habitat specialists with patchy distributions. This can make it difficult to determine the extent of a species range if there is strong sample selection bias (McPherson & Jetz, 2007; Seoane et al., 2005). Sample selection bias can often occur when numerous occurrence records are collected in areas more amenable to sampling, such as roadsides or areas with higher human population density (Fletcher & Fortin, 2018). Despite advancements in predictive algorithms and the pressing need for modeling rare species, there are few SDM studies addressing methods for modeling species with limited distribution data (Lomba et al., 2010; Mousikos et al., 2021).

Community science (sometimes called "citizen science") platforms such as iNaturalist have emerged as alternative sources for species occurrence data over the last decade (Gardiner et al., 2012; Mesaglio & Callaghan, 2021). Using iNaturalist, amateur naturalists can upload georeferenced photographs of a species observation. This platform provides large amounts of species distribution data, with over 126 million observations to date (https://www.inaturalist.org/observations). Although these datasets can be powerful, there is concern that these data are of lower quality than natural history collections (Gardiner et al., 2012). Correctly determining a species identification can be challenging for many nonexperts, even if they have some familiarity with the subject species (Silvertown et al., 2015); however, expert users on the platform can help by corroborating or correcting identifications. Geographic accuracy of observation points may be lower as well due to issues relating to privacy and errors, such as incorrectly uploading information from mobile devices (Suzuki-Ohno et al., 2017). Sample selection bias is another potential issue, with observations often skewed toward low-effort sampling sites, such as along roadsides and in areas with high population density (Armstrong, 2021). Despite concerns over quality, these datasets may have promise for improving SDMs when there is limited distributional data available for a species in natural history collections and for serving as independent datasets for model validation. This is especially promising since collection of new museum records for rare species may be limited by legal protections and ethical concerns. Therefore, further investigation into the utility of iNaturalist data for modeling the distribution of rare species is warranted.

In this paper, we used a combination of herbarium and iNaturalist records, as well as five different modeling techniques, to address the challenge of developing SDMs for data-limited rare species. We first evaluate the effectiveness of iNaturalist and natural history collection records as model training and evaluation data. To illustrate this approach, we develop SDMs for a test species, *Telesonix jamesii* (James' telesonix). *T. jamesii* is a rare, high-elevation plant endemic to the southern Rocky Mountains of Colorado and New Mexico, USA. Because methods for modeling rare species are limited in the literature, five different modeling techniques were used, including profile methods, statistical models, and machine learning algorithms. Using the best-performing modeling technique, we compare SDM performance using only herbarium data and

a combination of the herbarium and iNaturalist data. Models are validated using cross-validation as well as iNaturalist data as an external validation dataset. We compare model results and discuss their relevance toward conservation of this species and similar species.

METHODS

Study species

T. jamesii (Figure 1) is a rare species of Saxifragaceae that grows in rocky habitats and is regionally endemic to the Southern Rocky Mountains. There are 21 known populations of this species, extending from northern New Mexico to Rocky Mountain National Park in northern Colorado (Beatty et al., 2004; Figure 1). *T. jamesii* grows from montane to alpine life zones, typically on granite tors in dry, poor nutrient soils in areas with high exposure to wind and UV radiation (Ackerfield, 2015). These conditions are common above tree line, making the alpine a suitable zone for this species. However, these conditions are often not as common in the subalpine zones, with large areas of forests isolating suitable exposed habitat for *T. jamesii* (Gaier et al., 2023). This species is an ideal candidate for a

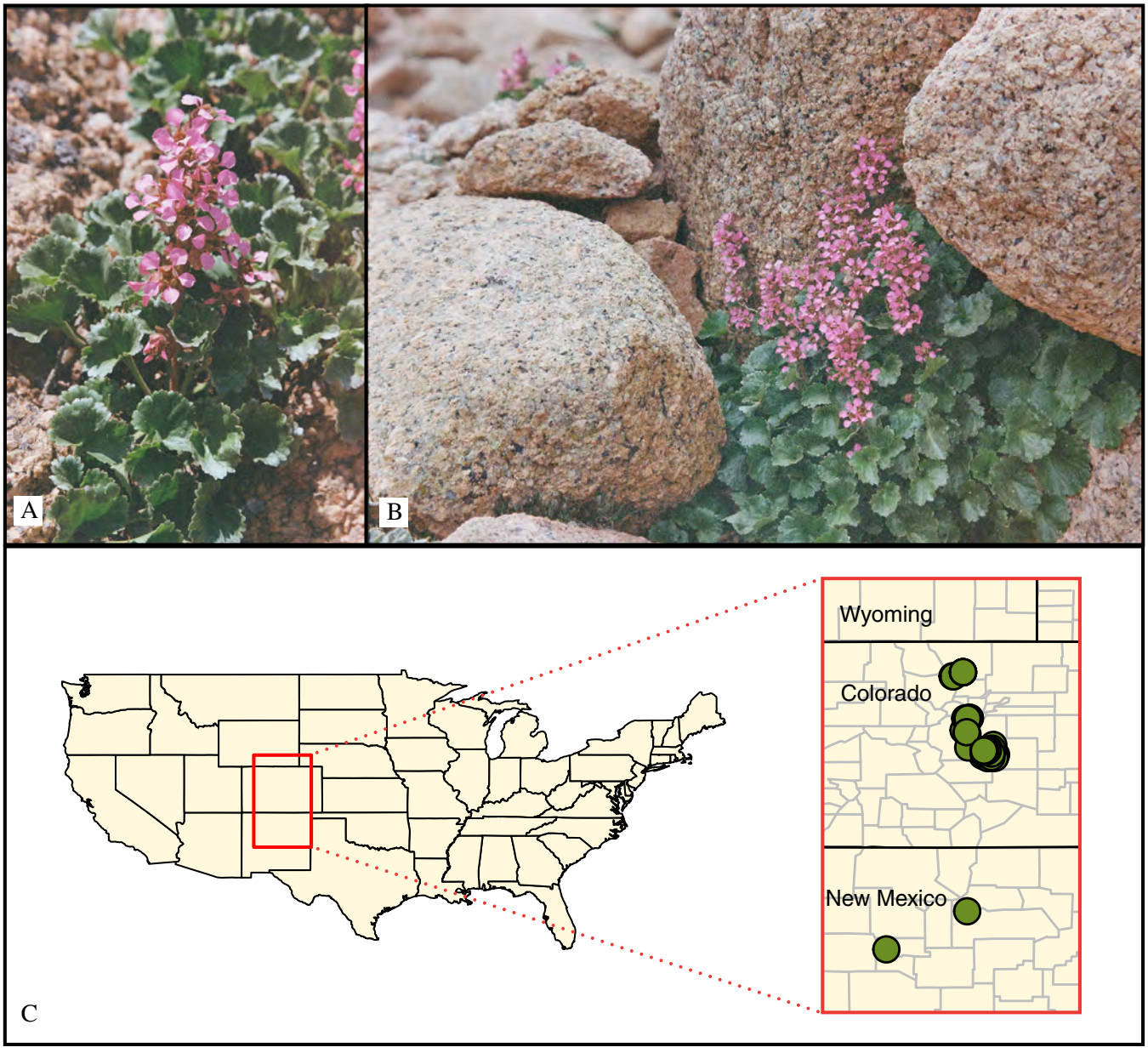

FIGURE 1 (A, B) *Telesonix jamesii* flowering in an exposed rocky habitat in the alpine of Pikes Peak, CO. Photos: K. Barthell. (C) The spatial extent considered for this study, encompassing the range of *T. jamesii* in the southern Rocky Mountains, USA, is indicated with red polygon. Herbarium records of *T. jamesii* occurrences are shown on the right.

community-science-based study on rare plants for several reasons. Although rare statewide, *T. jamesii* is abundant locally surrounding Pikes Peak and Rocky Mountain National Park, which are two highly visited wilderness areas in Colorado. The bright and charismatic flowers of this species make it unlikely to be passed up by keen iNaturalist users. Furthermore, these distinctive floral traits make it easier to validate observations on iNaturalist. This species is also threatened; it is ranked as S2 (imperiled in state because of rarity; Beatty et al., 2004) by the Colorado National Heritage Program. Thus, rigorous distribution modeling of *T. jamesii* could benefit conservation efforts for this species (Gaier et al., 2023).

Herbarium data

We used herbarium records of *T. jamesii* reported on SEINet (https://swbiodiversity.org/seinet/). There were initially 201 herbarium records of *T. jamesii*. *T. jamesii* and its only congener, *Telesonix heucheriformis*, have occasionally been treated as the same species in the past (Beatty et al., 2004). Distinct morphological and distributional differences are now considered necessary to separate these two species (Gornall & Bohm, 1985). *T. heucheriformis* has darker colored, shorter corollas and is found in Nevada, Utah, Wyoming, Montana, and South Dakota (Beatty et al., 2004). There were seven *T. jamesii* specimens collected in Nevada, Wyoming, and Montana, which we suspected were either identified incorrectly or never taxonomically revised (Appendix S1: Table S1). All of these specimens were in close geographic proximity to other *T. heucheriformis* occurrences, leading us to further suspect that they were not *T. jamesii*. We contacted the herbaria housing these specimens and requested either detailed photographs or confirmation of identification from herbarium staff to determine whether these specimens were *T. heucheriformis* (Appendix S1: Table S1). All seven specimens were confirmed to be *T. heucheriformis* and were eliminated from the dataset. We also eliminated any specimens without geocoordinates. Observations dated prior to 1970 were removed from the dataset as well. There were 32 specimens with geocoordinates dated prior to 1970. We chose 1970 as our cutoff date because we do not have climate data prior to 1970 (see *Environmental data*). To avoid pseudoreplication, we eliminated any observations from the same 900-m (30 arcseconds) cell as another. The resolution of our climate rasters is 900 m, which is the coarsest resolution of all of our environmental data (Table 1). In total, 30 herbarium specimen records remained (Figure 1; Appendix S1: Table S2).

TABLE 1 List of initial environmental predictors considered in our models.

Variable	Description
Worldclim,[a] 900 m	
Bio1	Annual mean temperature
Bio2	Mean diurnal range
Bio3	Isothermality
Bio4	Temperature seasonality
Bio5	Max temperature of the warmest month
Bio6	Min temperature of the coldest month
Bio7	Temperature annual range
Bio8	Mean temperature of the wettest quarter
Bio9	Mean temperature of the driest quarter
Bio10	Mean temperature of the warmest quarter
Bio11	Mean temperature of the coldest quarter
Bio12	Annual precipitation
Bio13	Precipitation of the wettest month
Bio14	Precipitation of the driest month
Bio15	Precipitation seasonality
Bio16	Precipitation of the wettest quarter
Bio17	Precipitation of the driest quarter
Bio18	Precipitation of the warmest quarter
Bio19	Precipitation of the coldest quarter
SRTM, 90 m	
Elevation	DEM
SoilGrids.org,[b] 250 m	
Cation	Cation exchange capacity of the soil (mmol/kg)
Nitrogen	Total soil nitrogen (cg/kg)
MRLC,[c] 90 m	
Barren	Presence or absence of barren landcover
Canopy	Percent canopy cover
Derived from DEM,[d] 90 m	
Slope	Slope between cells in DEM
Aspect	Orientation of slope measured clockwise from 0 to 360

Abbreviations: DEM, digital elevation model; SRTM, Shuttle Radar Topography Mission.
[a]https://www.worldclim.org/.
[b]https://www.isric.org/explore/soilgrids.
[c]Multi-Resolution Land Characteristics Consortium, https://www.mrlc.gov/.
[d]https://www.rdocumentation.org/packages/raster/versions/3.5-15/topics/terrain.

iNaturalist data

T. jamesii observations were downloaded directly from iNaturalist in January 2022. Because there is often error when users upload GPS information on their mobile devices,

we eliminated any observations that were placed in unlikely habitats, such as residential areas. We confirmed the identification of all observations using diagnostic morphological characteristics from user-uploaded images. Relative corolla lengths and coloration were examined to distinguish between *T. jamesii* and *T. heucheriformis* (Gornall & Bohm, 1985). We again thinned observations occurring in the same 900-m cell as other observations. In total, 29 iNaturalist records remained (Appendix S1: Table S3).

Environmental data

Predictor variables were derived from online sources (Table 1). These include 19 climatic variables from WORLDCLIM averaged between 1970 and 2000 (Hijmans et al., 2005; http://www.worldclim.org/bioclim.htm). We downloaded climate data at resolution of 30 arcseconds (900 m). Other variables considered were topographical features like elevation, slope, and aspect. We obtained a digital elevation model from the Shuttle Radar Topography Mission (Farr & Kobrick, 2000) at a resolution of 90 m. Slope and aspect were derived from a digital elevation model by taking the mean difference between the value of a cell and the values of its eight surrounding cells (Hijmans, 2022). Additional variables that are biologically relevant to *T. jamesii* were used. The life history strategy of *T. jamesii* promotes growth in open, rocky environments (Beatty et al., 2004; Gaier et al., 2023). As a proxy for this habitat type, land cover maps were downloaded from the National Land Cover Database (NLCD) and filtered to include only barren landcover. Barren landcover includes gravel, bare rock, and talus slopes. We incorporated barren landcover into the model as a binary covariate. Canopy cover was also downloaded from NLCD. Both canopy and landcover were downloaded at a resolution of 30 × 30 m and then aggregated to 90 × 90 m to aid computation time. Soil maps can be particularly appropriate for modeling edaphically limited species (Velazco et al., 2017). Maps of soil nitrogen content and cation exchange capacity were downloaded from SoilGrids (https://soilgrids.org/). This is a global database of estimated soil properties mapped at a 250 × 250 m spatial resolution.

Once all the raster layers were obtained, they were cropped to the model extent. We defined the extent as −108° W, −103° W, 34° N, and 42° N (Figure 1). This includes the entire Colorado Front Range as well as potential northern and southern habitats in Wyoming and New Mexico. Once projected over the study area, layers were resampled to a resolution of 90 × 90 m. We used bilinear interpolation to resample rasters that had resolutions coarser than 90 × 90 m.

Background data

To compensate for having no absence data (e.g., from systematic surveys), presence-only data can be contrasted with pseudoabsences, which are randomly drawn background points across the study extent (Fletcher & Fortin, 2018). These have been shown to produce more accurate SDMs than presence-only data on their own (Elith et al., 2005). Pseudoabsences are particularly appropriate for modeling rare endemics due to the higher probability that points selected will be true absences (Williams et al., 2009). We generated a different set of pseudoabsences for our two training datasets (only herbarium and herbarium with iNaturalist). Here we chose to generate 16 pseudoabsences for every true presence. Liu et al. (2019) found that for rare species, the number of background points needs to be up to 16 times greater than the number of true presences for model accuracy to reach an asymptote. We randomly generated pseudoabsences across the extent of the study area, with no points falling within 4.5 km of a true presence to help ensure that pseudoabsences did not fall on suitable habitat.

Variable selection

An important decision to make when fitting SDMs with limited distribution data is the selection and number of environmental variables (Guisan & Zimmermann, 2000). Overfitting a model with too many predictors can affect accuracy and predictive power; however, the inclusion of many variables may provide more informative models (Lomba et al., 2010). This trade-off makes variable selection a challenging decision when working with rare species. An effective method for thinning variables is to eliminate those that are highly correlated (Zuur et al., 2009). The variance inflation factor (VIF) measures how strongly each predictor can be explained by the rest of the predictors and is one of the most widely used methods for dealing with collinearity (Naimi et al., 2014). We eliminated the variables with the highest VIFs in a stepwise process until only variables with VIFs below three remained. A VIF of three is a conservative threshold for model collinearity (Zuur et al., 2009).

Model fitting

Recent developments in SDM packages have allowed modelers to use different predictive algorithms. Although it is promising that these modeling techniques can be easily applied to species occurrence data, it has been found that distinct modeling techniques can turn out

different results when calibrated on the same species (Broennimann et al., 2007). Therefore, assessing which modeling technique works best for *T. jamesii* is necessary before further investigating the more specific uses of iNaturalist data. To do so, we first combined all occurrence data and fitted an array of models (Figure 2).

To capture the full range of modeling techniques, we considered profile methods, statistical models, and machine learning algorithms. Profile methods relate environmental variability at presence locations to background data across the study extent using similarity-based measures (Fletcher & Fortin, 2018). We implemented a BIOCLIM model, which is a commonly used profile method and the first SDM package (Booth, 2018). Statistical models used in SDMs are commonly variations on linear models. We used generalized linear models (GLMs) and generalized additive models (GAMs). Although GLMs are widely used for distribution modeling, a major concern is that they fail to capture nonlinear relationships between species with predictor variables (Elith et al., 2006). GAMs offer an alternative, using splines to accommodate nonlinearity in response functions (Fletcher & Fortin, 2018). No interaction terms or quadratic terms were considered for statistical models. Finally, we implemented two machine learning algorithms: random forest (RF) and Maxent. RF works by growing a suite of regression trees that bootstrap the original data. The outcome of each bootstrap informs the algorithm on how to fit the model (Fletcher & Fortin, 2018). The number of explanatory variables (mtry) that are sampled for each tree can be adjusted manually to minimize predictive error (Fletcher & Fortin, 2018). After tuning our RF model, we set mtry as 1. Maxent is one of the most widespread and routine algorithms for SDMs in scientific studies and applied modeling (Lissovsky & Dudov, 2021). These models are based on the principle of maximum entropy, which states that the most uniform distribution is the best approximation of an unknown distribution (Phillips et al., 2006). The regularization parameter was tuned manually to reduce error. Here, we specified our beta multiplier as 1 in our Maxent model.

Model evaluation and selection

We used K-fold cross-validation to evaluate the performance of this first suite of models. K-fold validation splits the training data into K equal-sized parts (Naimi et al., 2014). Each K-fold is used as model testing data, and the other K-1 folds are used as training data. We used five folds for evaluating our data. There are many statistical approaches for evaluating model performance (Fielding & Bell, 1997). K-fold cross-validation is an appropriate method for this study because, although it is not truly independent, each validation fold replicates an

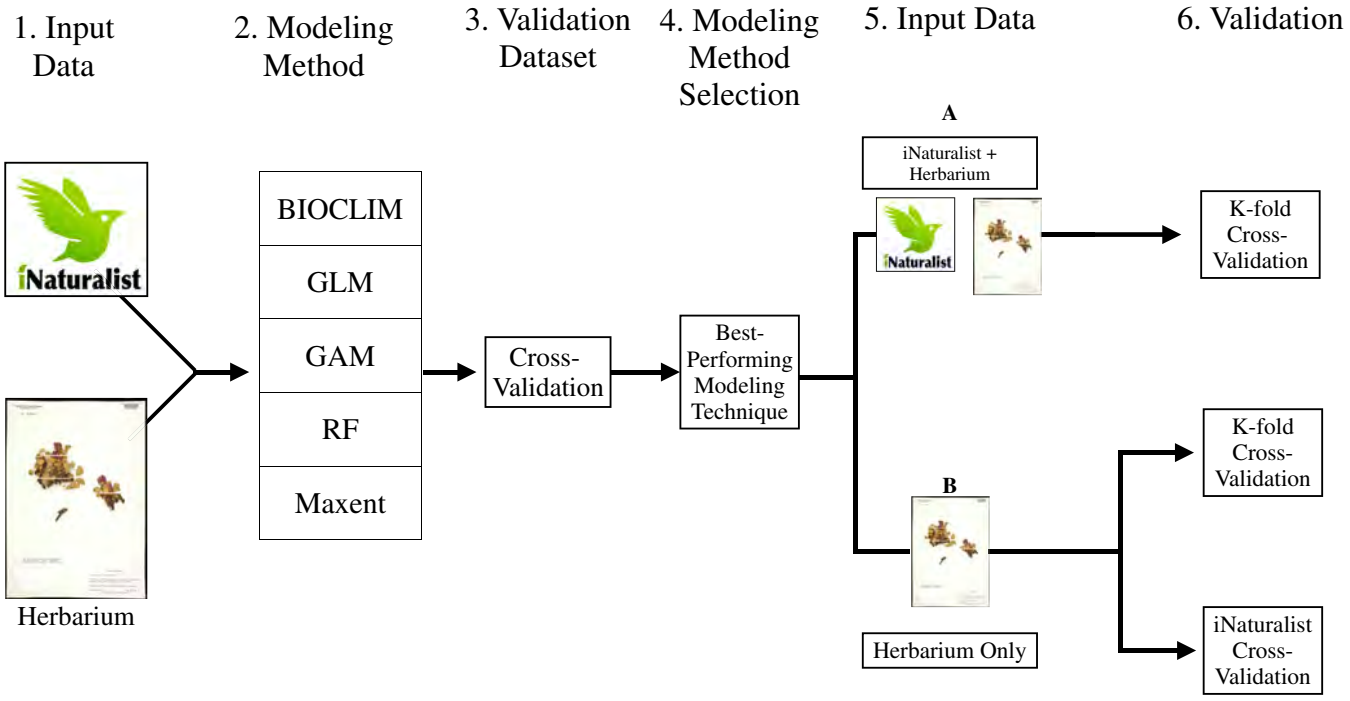

FIGURE 2 Overview of successive steps in our model-building process. Models are either trained/tested with data from iNaturalist, herbaria, or the two combined. Model evaluation metrics are used to select the best-performing algorithm at step 4. GAM, generalized additive model; GLM, generalized linear model; RF, random forest.

independent dataset (Pearson, 2010). This promotes consistency between our evaluations since we are examining both internally validated models and externally validated models. Here, we evaluated performances using both a threshold-independent metric (area under curve, AUC) and a threshold-dependent metric from a confusion matrix, true skill statistic (TSS; Lawson et al., 2014). We also report sensitivity and specificity from model evaluations.

To determine where each model predicts a presence or absence, the distribution probabilities for each model were converted into a binary map of species presence and absence by applying a threshold value (Appendix S1: Figure S1; Bagaria et al., 2021). The threshold value for each model was calculated based on the maximization of sum of specificity and sensitivity.

After selecting the best-performing modeling technique based on evaluation metrics, we developed two new models, each with different training data (Figure 2). Our first model was trained using both herbarium and iNaturalist data and evaluated once again using K-fold validation. Our second model was trained using only herbarium data and validated with training data and again with only iNaturalist data. This second evaluation with just iNaturalist data is the only instance in this study where a model is externally evaluated with a truly independent dataset.

All analyses were performed in R version 4.0.2 (R Core Team, 2022). We used the raster package version 3.3-15 for analyses, visualizations, and manipulations of all raster layers, as well as generating background points (Hijmans, 2022). The usdm package version 1.1-18 was used for thinning variables and diagnosing collinearity (Naimi et al., 2014). For training and testing our SDMs, we used the dismo package version 1.3-5 (Hijmans et al., 2021), the PresenceAbsence package version 1.1.10 (Freeman & Moisen, 2008), the randomForest package version 4.7-1 (Liaw & Wiener, 2002), and the glmnet package version 4.1-1.3 (Friedman et al., 2010). The sp package version 1.4-6 (Pebesma & Bivand, 2005), the rgdal package version 1.5-28 (Bivand et al., 2021), and the dplyr package version 1.0.8 (Wickham et al., 2022) were used for various data formatting tasks.

TABLE 2 Variance inflation factors (VIFs) of environmental variables remaining after diagnosing collinearity problems.

Variables	Description	VIF
Bio2	Mean diurnal range	1.261749
Bio4	Temperature seasonality	1.924307
Bio6	Min temperature of the coldest month	2.972027
Bio9	Mean temperature of the driest quarter	1.308481
Bio18	Precipitation of the warmest quarter	1.503391
Nitrogen	Cation exchange capacity of the soil (mmol/kg)	2.999971
Cation	Total soil nitrogen (cg/kg)	1.919629
Slope	Slope between cells in DEM	1.878356
Aspect	Orientation of slope measured clockwise from 0 to 360	1.014600
Canopy	Percent canopy cover	1.277840
Barren	Presence or absence of barren landcover	1.015801

Note: The VIFs of all variables were taken, and we eliminated the variable with the highest VIF. We then took the VIFs once again and continued this process until all variables had VIFs below three.
Abbreviation: DEM, digital elevation model.

RESULTS

Model performance and selection

We ran our initial suite of models using a reduced set of environmental variables. Of our initial set of variables, only 11 remained after meeting our VIF threshold (Table 2). The predictive maps for all five initial models (BIOCLIM, GLM, GAM, Maxent, and RF) showed similar patterns of low-probability habitat across most of the study extent, with areas of higher probability surrounding Pike National Forest and other small pockets across the Front Range (Figure 3). Despite two herbarium occurrences in New Mexico, none of the models predicted high-probability habitat in the southern end of the study area. Interestingly, many of our models showed suitable habitat in the Sangre de Cristo Mountains near the Colorado–New Mexico border, where there are no known occurrences. All models had AUC values greater than 0.85, which indicates high predictive accuracy (Lomba et al., 2010). The optimal threshold for presence varied across models (Figure 3). GLM recorded the lowest optimal threshold (0.079). The BIOCLIM model predicted the least amount of high-probability habitat compared to the other models (Figure 3). This algorithm also displayed both the lowest AUC and TSS values (Table 3), despite displaying the second highest specificity value (0.964). This is likely due to our BIOCLIM model showing the lowest sensitivity of any model (0.750). BIOCLIM and GAM recorded the highest optimal thresholds for detecting a presence (Table 3). The projections for these models show more high-probability habitat compared to the other model projections (Figure 3). This would call for a higher threshold so that the sum of specificity and sensitivity is maximized. Although models varied in their sensitivity, specificity was high across all models (>0.90),

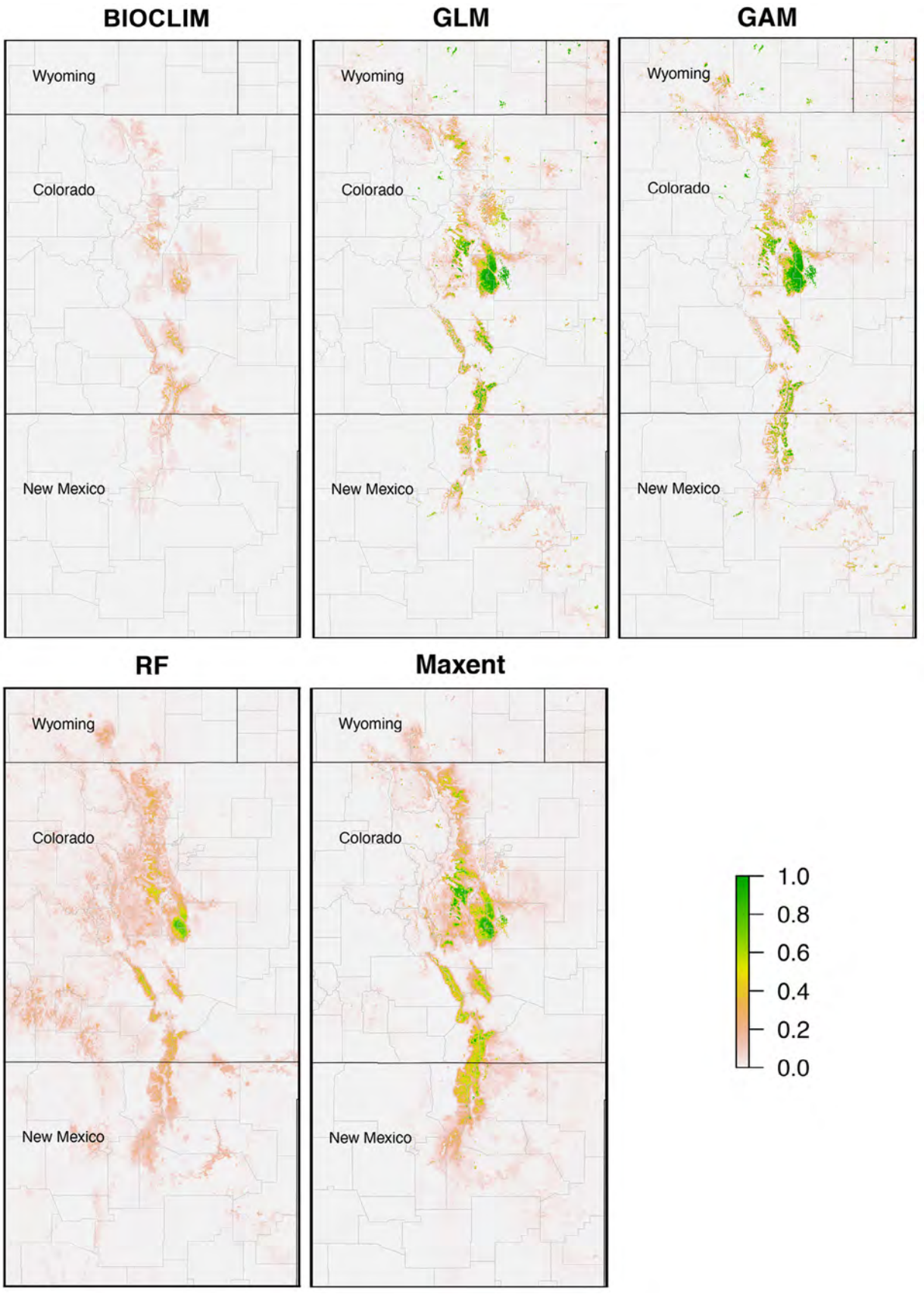

FIGURE 3 Predictive maps of the distribution of *Telesonix jamesii* across the study region generated by each modeling technique. A scale indicating probability of occurrence is shown to the right of each map. State and county boundaries are included for geographic reference. GAM, generalized additive model; GLM, generalized linear model; RF, random forest.

indicating that models were overall more successful at predicting absences than presences.

Of the five different modeling techniques implemented, RF had the highest TSS and AUC values when cross-validated (Table 3). With an AUC of 0.981, the initial RF model can be considered an excellently fitted model (Elith & Leathwick, 2007). Thus, the RF modeling algorithm was utilized for the continuation of the study.

The ranked importance of each predictor considered in our RF model is summarized in Figure 4. Of our 11 predictors, Bio9 (mean temperature of the driest quarter) and Bio18 (precipitation of the warmest quarter) were scored the highest by RF by a wide margin. Species response curves (Appendix S1: Figure S2) to the most important predictors (Bio9 and Bio18) reveal that *T. jamesii* is preferably distributed in areas with high summer rainfall and cold temperatures during the quarter with the lowest precipitation, which is winter in western North America (Krause et al., 2015; Schwinning et al., 2005). Cation exchange, Bio4, and Bio2 were all ranked intermediately (Figure 4). Despite what is known about the habitat preferences of *T. jamesii* (Gaier et al., 2023), canopy and barren landcover were among the least important variables for predicting its distribution. Overall, these results show the dominance of variables related to precipitation regimes in determining the range of our test species.

TABLE 3 Model metrics for the first suite of models.

Model	AUC	TSS	Sens	Spec	Threshold
BIOCLIM	0.868	0.714	0.75	0.964	0.151
GLM	0.96	0.889	0.956	0.933	0.079
GAM	0.901	0.733	0.754	0.979	0.445
RF	0.981	0.896	0.964	0.932	0.11
Maxent	0.975	0.881	0.912	0.941	0.0895

Abbreviations: AUC, area under curve; GAM, generalized additive model; GLM, generalized linear model; RF, random forest; Sens, sensitivity; Spec, specificity; TSS, true skill statistic.

iNaturalist as an evaluation dataset

RF models performed slightly differently depending on how iNaturalist data were used (Table 4). The model built using only herbarium data recorded slightly higher test statistics when cross-validated with iNaturalist data (AUC = 1, TSS = 1) compared to when it was cross-validated with herbarium data (AUC = 0.988, TSS = 0.935). Our model using herbarium and iNaturalist data together as both training and testing data displayed the lowest predictive accuracy between our three RF models (AUC = 0.981, TSS = 0.896). It should be noted that all models had AUC values greater than 0.95 and TSS values above 0.85, which suggests that all models had an excellent fit (Allouche et al., 2006; Guisan et al., 2017).

DISCUSSION

Rare species have long been considered a challenge to model due to severe limitations of occurrence data (Lomba et al., 2010). Scarce datasets not only limit the geographic breadth of modeling procedures but also hamper the number of predictor variables that can be incorporated into models (Guisan & Thuiller, 2005; Walther et al., 2007).

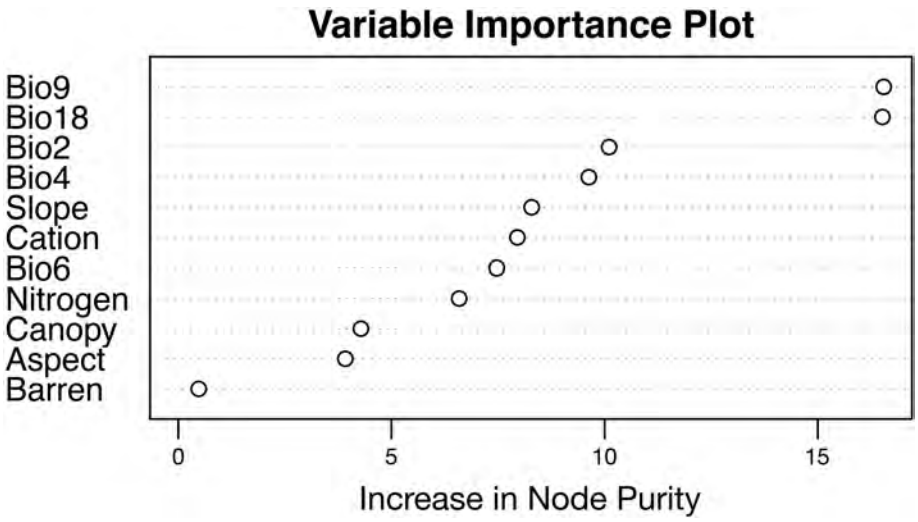

FIGURE 4 Ranked importance of each predictor variable considered by random forest. Importance is measured by the increase in node purity. Higher values indicate that the explanatory variable is an important predictor for *Telesonix jamesii* distribution. See Table 2 for an explanation of variable abbreviations.

TABLE 4 Test statistics for the second batch of random forest models based on which data are used for training and which data are used for testing.

Training data	Testing data	TSS	Sens	Spec	AUC
Herbaria	Herbaria	0.935	1.00	0.935	0.988
Herbaria	iNaturalist	1.00	1.00	1.00	1.00
Herbaria and iNaturalist	Herbaria and iNaturalist	0.896	0.96	0.932	0.981

Abbreviations: AUC, area under curve; Sens, sensitivity; Spec, specificity; TSS, true skill statistic.

This then further limits the possibility of obtaining a fuller understanding of this species' niche, despite the crucial need to investigate species–environmental relationships of rare and threatened species (Farnsworth & Ogurcak, 2006). Therefore, procedures for dealing with this "rare species modeling paradox" must be explored (Lomba et al., 2010). In our approach to the "rare species modeling paradox," we incorporated iNaturalist data into our dataset along with the more traditional herbarium data. Although iNaturalist records are typically thought to be of lower quality compared to museum collection records (Gardiner et al., 2012), we found more erroneous observations in our herbarium data. This is surprising, especially given that herbarium collections are considered the gold standard for plant species occurrence records. Additionally, these erroneous observations were only flagged because we were familiar with the biology of our study species, which may not always be the case in other applied modeling scenarios. We omitted only two iNaturalist observations from our dataset due to inaccurate geocoordinates. One observation was from inside the Pikes Peak gift shop, and the other was in a Colorado Springs residential area. These were both likely the result of mobile device error in assigning coordinates to the observations. The herbarium data for *T. jamesii* contained seven erroneous presences, all of which were the result of incorrect taxonomic classification. Much more effort was then required to confidently omit these specimens from our dataset compared to the iNaturalist data. It is however important to note that a vast number of plants require direct observation with a hand lens or dissecting scope to be correctly identified, which cannot be achieved with most iNaturalist photos. Our data integration approach would be best applied to species with traits that make filtering out erroneous iNaturalist observations possible. The fact that *T. jamesii* has distinct flowers and is a habitat specialist makes that process relatively easy; however, mistakes might be more difficult to catch if the species has minute diagnostic characteristic or a wider niche breadth. It is also important to consider that this species had a similar number of herbarium and iNaturalist records (herbarium = 30, iNaturalist = 29). The approach shown in this paper may not be as promising in cases where the number of herbarium records greatly outweighs the number of iNaturalist records; however, instances where there are much more iNaturalist records would likely benefit from our approach.

From our framework, we generated an initial suite of five commonly used modeling algorithms. The projections for each of these models appear qualitatively different from one another. GAM, for instance, predicted more total area of high-quality habitat, whereas BIOCLIM predicted less. These different projections could have biological significance for this species. For example, our GAM and Maxent model predict a lot more habitats outside the known range of this species, which could indicate that this species is limited by dispersal. There appears to be suitable habitat elsewhere, but *T. jamesii* may not have the ability to establish in these favorable areas. A more conservative projection such as RF might indicate that *T. jamesii* is more constrained by the environmental factors at the locations where it is already found. Despite these differences, the evaluation metrics between all models were relatively high. This is likely due to the variation in optimal threshold values between models. Another consideration to make when selecting a modeling algorithm is computation time and interpretability. All models ran relatively quickly, with GLM being the quickest and longer run times for BIOCLIM and GAM. Although computation time was not an issue in this study, it could present challenges when modeling species with larger ranges and more presences. Ease of interpretability can vary between algorithms. It may be more difficult to infer the direct relationship between probability of occurrence and environmental variables when using nonparametric algorithms such as RF and Maxent, whereas statistical approaches like GLM will give us a more direct inference on these relationships. Ensemble models can sometimes offer an alternative to using a single modeling algorithm when performance varies between models (Araújo & New, 2007). Because all models performed well and the goal of this study was to evaluate the efficacy of different models for a rare species, we chose not to consider an ensemble model. Additionally, although ensemble models can sometimes be more accurate than single model predictions (Marmion et al., 2009),

it is important to consider that the models we use are fundamentally predicting different things. For example, an envelope model, such as BIOCLIM, predicts environmental similarity while a statistical method, such as GLM, predicts the probability of occurrence (Fletcher & Fortin, 2018). We found that our RF model had the best fit of all five models (Figure 5). This was surprising because Maxent is a more widely used modeling algorithm for SDMs (Fletcher & Fortin, 2018) and has often been treated as the default method for modelers due to its prevalence (Fourcade et al., 2014). We show that it is informative to investigate a number of different modeling techniques when dealing with rare species.

Proceeding with RF, we found that our model performed the best when using herbaria data as training data and iNaturalist data as testing data, though all combinations tested performed very well. External independent datasets often provide more robust evaluations of presence-only models (Araújo et al., 2005). Since these datasets are difficult for modelers to obtain, the utility of iNaturalist data for validation shown in this study is promising. However, there were drawbacks to using iNaturalist for validating our SDM for *T. jamesii*. Collections from herbaria ranged from northern Colorado through New Mexico. iNaturalist data did not span that range, with all of the observations derived from Colorado. Even though our herbaria-only model was accurate at predicting the iNaturalist occurrences, the iNaturalist occurrences did not represent the entire distribution of this species. Similarly, the herbaria-only model was not as accurate when cross-validated compared to being externally validated. This suggests that it had difficulty predicting suitable habitat across the entire study extent outside of the few prominent populations in Colorado. Because many narrow endemics follow similar patchy distributions across their ranges (Kruckeberg & Rabinowitz, 1985), more localized

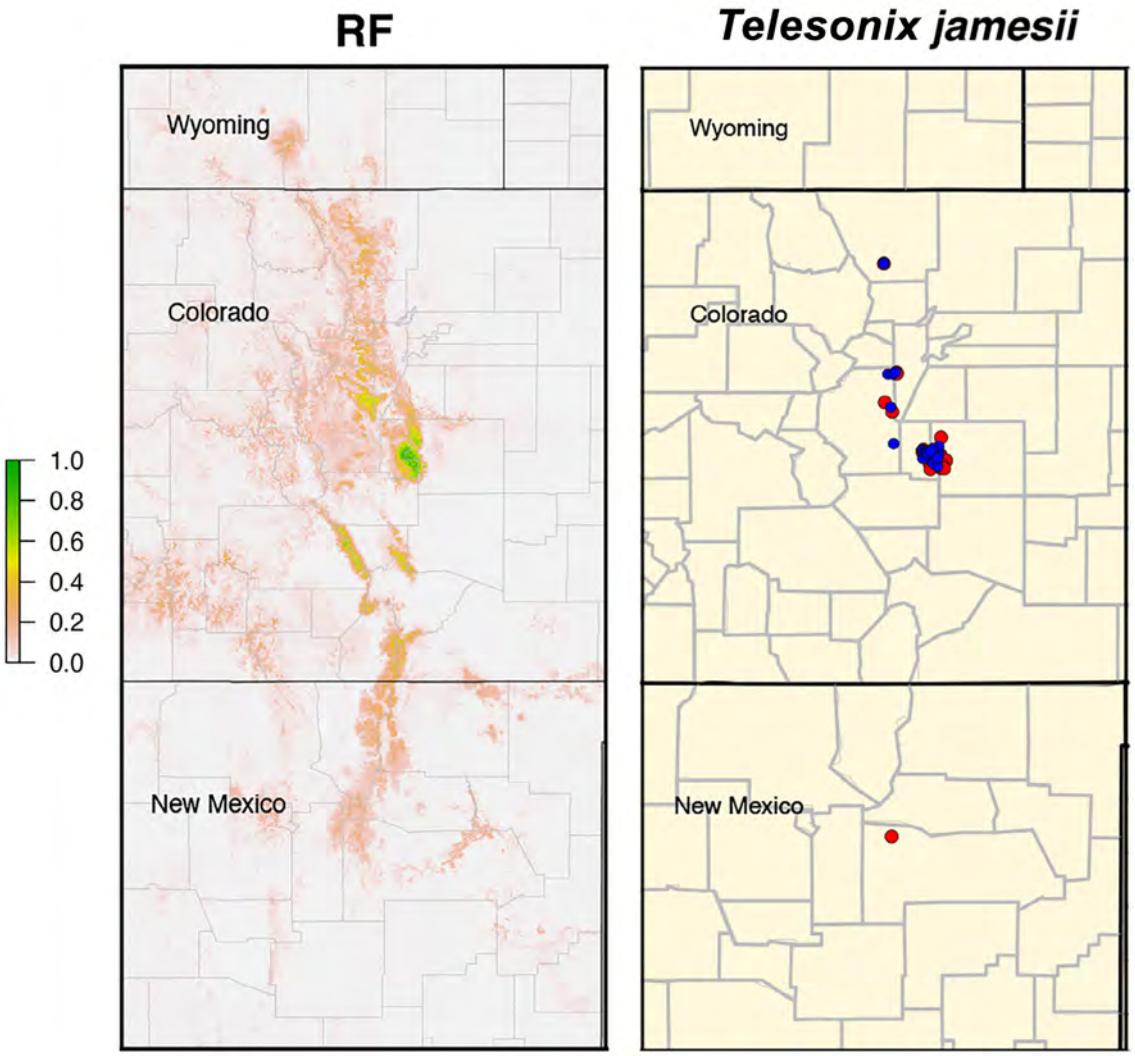

FIGURE 5 Best-performing random forest (RF) model (left) showed in contrast with occurrences of herbarium and iNaturalist records (right). Red dots indicate presence from herbarium dataset and blue dots indicate presence from iNaturalist dataset.

validation datasets like the one used in this study may not be as promising as their test statistics suggest.

An important factor to consider when training SDMs is scale (Fletcher & Fortin, 2018). Although some of our predictor variables, such as elevation and landcover, were mapped at finer resolutions, all of our climatic predictors were at a grain size of 900 m. Variation in climate is generally not as drastic across most landscapes, but the steep gradients of alpine and subalpine environments bring about changes in microclimates at much smaller scales (Wershow & DeChaine, 2018). This variation is not likely represented in the climatic predictors used in these models. Another caveat is that although *T. jamesii* follows a distribution pattern similar to many other rare endemics, most of the known subpopulations are centered around Pikes Peak and Rocky Mountain National Park, which are among the most highly trafficked wilderness areas in Colorado. This means that *T. jamesii* may have more abundant iNaturalist data compared to other species with patchy distributions. Therefore, our approach may be less effective for species that are not distributed around popular destinations.

Another important consideration when modeling rare species is distinguishing whether the small number of records is because observations have been made at a subset of populations or if few records are available because the species only occurs in a few locations to begin with. In the latter case, the data may provide an accurate representation of a species distribution despite having limited records. It is possible that the records available for *T. jamesii*, while few, accurately represent its distribution, but that is difficult to truly discern. In either case, we cannot know for certain the true distribution of this species, and we will always be biased by data we have in any model. Although cleaning the data spatially and temporally as we have done here can help us reduce biases, these concerns may very well persist in the remaining records. Specifically, it is difficult to separate the effects of high sampling efforts in the highly visited areas of Pikes Peak and Rocky Mountain National Park from the true distribution of *T. jamesii*. SDMs can also suggest areas of suitable habitat where unrecorded populations might occur. Our best-performing model (RF) predicted suitable habitat in areas with records and also in other areas where no occurrences have been recorded, such as in southern Colorado. Targeted surveys of this region could potentially identify new populations. Expert evaluation of model projections can often aid in validating SDMs (Fourcade, 2016; Gastón et al., 2014); however, expert knowledge may also be bound to the same distributional records that are already used in SDMs. We shared our results with botany experts at several institutions in Colorado: the Colorado Natural Heritage Program, the University of Colorado Herbarium, and the Colorado Native Plant Society—none of which knew of any anecdotal occurrences in the Sangre de Cristo Mountains. Additionally, none reported knowledge of any occurrences outside the areas of our known records. It is also possible that an area with suitable habitat for *T. jamesii* may not be occupied due to dispersal limitations preventing the colonization of these typically isolated mountain habitats. In this case, areas like the Sangre de Cristos may be a potential refuge for this species if land managers or conservation practitioners were ever to transplant individuals from other populations.

There is a wide potential usefulness for the data integration approach shown in this study toward the conservation and management of rare species. We have shown that community science data can be reliable and substantially increase the number of usable records leveraged for modeling distributions. Both museum records and iNaturalist records require examination to determine taxonomic and location reliability. The use of iNaturalist data improved model fits only slightly. Choice of modeling algorithm showed more variation in our results than choice of data source. Much of the information needed to accurately model *T. jamesii* distribution was already captured in the herbarium data. We therefore speculate that this framework may be more useful for a species with more iNaturalist observations in novel habitats. It is important to consider that this is a species-specific study, and greater insights could be gained through a multispecies approach. A potential next step would be to evaluate how this data integration approach differs between species displaying different patterns of rarity (Rabinowitz, 1981). Notwithstanding these caveats, the information obtained from our model projections can aid in the conservation of *T. jamesii* to support future targeted surveys (Williams et al., 2009), help identify populations most at risk, and predict how distributions may be affected by future climate change (Franklin, 2013). This study can serve as groundwork for future SDM studies of species with similar data limitations.

ACKNOWLEDGMENTS

We thank Dan Doak and Erin Manzitto-Tripp for helpful feedback on this manuscript; Ryan Langendorf and Dan Doak for assistance with computing resources; and Grace Kostel, Mark Gabel, Dan Potter, Ernie Nelson, Jerry Tiehm, Mitchell McGlaughlin, and Bob Dorn for helping us revise specimens in question across intermountain herbaria. We also thank the Colorado Natural Heritage Program and the Colorado Native Plant Society for letting us share and corroborate our results. This research was funded by the National Science Foundation (DEB award 2102974), a National Geographic Explorer Grant, the

Colorado Native Plant Society, and the Department of Ecology and Evolutionary Biology at the University of Colorado, Boulder.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Code and data (Gaier, 2023) are available from Figshare: https://figshare.com/articles/dataset/T_jamesii_SDM_files/21860205.

ORCID

Andrew G. Gaier https://orcid.org/0000-0002-5296-0439
Julian Resasco https://orcid.org/0000-0003-1605-3038

REFERENCES

Ackerfield, J. 2015. *Flora of Colorado*, 818. Fort Worth, TX: BRIT Press.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS)." *Journal of Applied Ecology* 43(6): 1223–32.

Araújo, M. B., and M. New. 2007. "Ensemble Forecasting of Species Distributions." *Trends in Ecology & Evolution* 22: 42–7.

Araújo, M. B., R. G. Pearson, W. Thuiller, and M. Erhard. 2005. "Validation of Species–Climate Impact Models under Climate Change." *Global Change Biology* 11(9): 1504–13.

Armstrong, Z. N. 2021. "Modeling Distributions of *Cantharellus formosus* Using Natural History and Citizen Science Data." (Order No. 28582707). ProQuest Dissertations & Theses A&I; ProQuest Dissertations & Theses Global.

Bagaria, P., A. Thapa, L. K. Sharma, B. D. Joshi, H. Singh, C. M. Sharma, J. Sarma, M. Thakur, and K. Chandra. 2021. "Distribution Modelling and Climate Change Risk Assessment Strategy for Rare Himalayan Galliformes Species Using Archetypal Data Abundant Cohorts for Adaptation Planning." *Climate Risk Management* 31: 100264.

Beatty, B. L., W. F. Jennings, and R. C. Rawlinson. 2004. *Telesonix jamesii (Torr.) Raf. (James' Telesonix): A Technical Conservation Assessment*. Lakewood, CO: USDA Forest Service, Rocky Mountain Region.

Bivand, R., T. Keitt, and B. Rowlingson. 2021. "rgdal: Bindings for the 'Geospatial' Data Abstraction Library." R Package Version 1.5-28. https://CRAN.R-project.org/package=rgdal.

Booth, T. H. 2018. "Why Understanding the Pioneering and Continuing Contributions of BIOCLIM to Species Distribution Modelling Is Important." *Austral Ecology* 43(8): 852–60.

Breiner, F. T., A. Guisan, A. Bergamini, and M. P. Nobis. 2015. "Overcoming Limitations of Modelling Rare Species by Using Ensembles of Small Models." *Methods in Ecology and Evolution* 6(10): 1210–8.

Broennimann, O., U. A. Treier, H. Müller-Schärer, W. Thuiller, A. T. Peterson, and A. Guisan. 2007. "Evidence of Climatic Niche Shift during Biological Invasion." *Ecology Letters* 10(8): 701–9.

Chen, I. C., J. K. Hill, R. Ohlemüller, D. B. Roy, and C. D. Thomas. 2011. "Rapid Range Shifts of Species Associated with High Levels of Climate Warming." *Science* 333(6045): 1024–6.

Elith, J., S. Ferrier, F. Huettmann, and J. Leathwick. 2005. "The Evaluation Strip: A New and Robust Method for Plotting Predicted Responses from Species Distribution Models." *Ecological Modelling* 186(3): 280–9.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, et al. 2006. "Novel Methods Improve Prediction of Species' Distributions from Occurrence Data." *Ecography* 29(2): 129–51.

Elith, J., and J. Leathwick. 2007. "Predicting Species Distributions from Museum and Herbarium Records Using Multiresponse Models Fitted with Multivariate Adaptive Regression Splines." *Diversity and Distributions* 13(3): 265–75.

Farnsworth, E. J., and D. E. Ogurcak. 2006. "Biogeography and Decline of Rare Plants in New England: Historical Evidence and Contemporary Monitoring." *Ecological Applications* 16(4): 1327–37.

Farr, T. G., and M. Kobrick. 2000. "Shuttle Radar Topography Mission Produces a Wealth of Data." *Eos, Transactions American Geophysical Union* 81(48): 583–5.

Fielding, A. H., and J. F. Bell. 1997. "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models." *Environmental Conservation* 24(1): 38–49.

Fletcher, R., and M. Fortin. 2018. *Spatial Ecology and Conservation Modeling*, 523. New York: Springer International Publishing.

Fois, M., G. Bacchetta, A. Cuena-Lombraña, D. Cogoni, M. S. Pinna, E. Sulis, and G. Fenu. 2018. "Using Extinctions in Species Distribution Models to Evaluate and Predict Threats: A Contribution to Plant Conservation Planning on the Island of Sardinia." *Environmental Conservation* 45(1): 11–9.

Fourcade, Y. 2016. "Comparing Species Distributions Modelled from Occurrence Data and from Expert-Based Range Maps. Implication for Predicting Range Shifts with Climate Change." *Ecological Informatics* 36: 8–14.

Fourcade, Y., J. O. Engler, D. Rödder, and J. Secondi. 2014. "Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias." *PLoS One* 9(5): e97122.

Franklin, J. 2013. "Species Distribution Models in Conservation Biogeography: Developments and Challenges." *Diversity and Distributions* 19: 1217–23.

Freeman, E. A., and G. Moisen. 2008. "PresenceAbsence: An R Package for Presence Absence Analysis." *Journal of Statistical Software* 23: 1–31.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1): 1–22.

Gaier, A. 2023. "*T. jamesii* SDM Files." Figshare. Dataset. https://doi.org/10.6084/m9.figshare.21860205.v1.

Gaier, A. G., E. Manzitto-Tripp, and J. Resasco. 2023. "Floral Visitors of a Colorado Endemic Chasmophyte, *Telesonix jamesii* (Saxifragaceae)." *Western North American Naturalist*: in press.

Gardiner, M. M., L. L. Allee, P. M. Brown, J. E. Losey, H. E. Roy, and R. R. Smyth. 2012. "Lessons from Lady Beetles: Accuracy of Monitoring Data from US and UK Citizen-Science Programs." *Frontiers in Ecology and the Environment* 10(9): 471–6.

Gastón, A., J. I. García-Viñas, A. J. Bravo-Fernández, C. López-Leiva, J. A. Oliet, S. Roig, and R. Serrada. 2014. "Species Distribution Models Applied to Plant Species

Selection in Forest Restoration: Are Model Predictions Comparable to Expert Opinion?" *New Forests* 45(5): 641–53.

Gornall, R. J., and B. A. Bohm. 1985. "A Monograph of Boykinia, Peltoboykinia, Bolandra and Suksdorfia (Saxifragaceae)." *Botanical Journal of the Linnean Society* 90(1): 1–71.

Guisan, A., and W. Thuiller. 2005. "Predicting Species Distribution: Offering More than Simple Habitat Models." *Ecology Letters* 8(9): 993–1009.

Guisan, A., W. Thuiller, and N. E. Zimmermann. 2017. *Habitat Suitability and Distribution Models: With Applications in R, Ecology, Biodiversity and Conservation*. Chicago, IL: The University of Chicago Press.

Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. Tulloch, T. J. Regan, et al. 2013. "Predicting Species Distributions for Conservation Decisions." *Ecology Letters* 16(12): 1424–35.

Guisan, A., and N. E. Zimmermann. 2000. "Predictive Habitat Distribution Models in Ecology." *Ecological Modelling* 135(2–3): 147–86.

Hijmans, R. 2022. "raster: Geographic Data Analysis and Modeling." R Package Version 3.5-15. https://CRAN.R-project.org/package=raster.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology: A Journal of the Royal Meteorological Society* 25(15): 1965–78.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith. 2021. "dismo: Species Distribution Modeling." R Package Version 1.3-5. https://CRAN.R-project.org/package=dismo.

Imperio, S., R. Bionda, R. Viterbi, and A. Provenzale. 2013. "Climate Change and Human Disturbance Can Lead to Local Extinction of Alpine Rock Ptarmigan: New Insight from the Western Italian Alps." *PLoS One* 8(11): e81598.

Koch, R., J. S. Almeida-Cortez, and B. Kleinschmit. 2017. "Revealing Areas of High Nature Conservation Importance in a Seasonally Dry Tropical Forest in Brazil: Combination of Modelled Plant Diversity Hot Spots and Threat Patterns." *Journal for Nature Conservation* 35: 24–39.

Krause, C. M., N. S. Cobb, and D. D. Pennington. 2015. "Range Shifts under Future Scenarios of Climate Change: Dispersal Ability Matters for Colorado Plateau Endemic Plants." *Natural Areas Journal* 35(3): 428–38.

Kruckeberg, A. R., and D. Rabinowitz. 1985. "Biological Aspects of Endemism in Higher Plants." *Annual Review of Ecology and Systematics* 16(1): 447–79.

Lavergne, S., W. Thuiller, J. Molina, and M. Debussche. 2005. "Environmental and Human Factors Influencing Rare Plant Local Occurrence, Extinction and Persistence: A 115-Year Study in the Mediterranean Region." *Journal of Biogeography* 32(5): 799–811.

Lawson, C. R., J. A. Hodgson, R. J. Wilson, and S. A. Richards. 2014. "Prevalence, Thresholds and the Performance of Presence–Absence Models." *Methods in Ecology and Evolution* 5(1): 54–64.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3): 18–22.

Lissovsky, A. A., and S. V. Dudov. 2021. "Species-Distribution Modeling: Advantages and Limitations of Its Application. 2. MaxEnt." *Biology Bulletin Reviews* 11(3): 265–75.

Liu, C., G. Newell, and M. White. 2019. "The Effect of Sample Size on the Accuracy of Species Distribution Models: Considering both Presences and Pseudo-Absences or Background Sites." *Ecography* 42(3): 535–48.

Lomba, A., L. Pellissier, C. Randin, J. Vicente, F. Moreira, J. Honrado, and A. Guisan. 2010. "Overcoming the Rare Species Modelling Paradox: A Novel Hierarchical Framework Applied to an Iberian Endemic Plant." *Biological Conservation* 143(11): 2647–57.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. "Evaluation of Consensus Methods in Predictive Species Distribution Modelling." *Diversity and Distributions* 15(1): 59–69.

McPherson, J. M., and W. Jetz. 2007. "Effects of Species' Ecology on the Accuracy of Distribution Models." *Ecography* 30: 135–51.

Mesaglio, T., and C. T. Callaghan. 2021. "An Overview of the History, Current Contributions and Future Outlook of iNaturalist in Australia." *Wildlife Research* 48: 289.

Mousikos, A., P. Manolaki, N. Knez, and I. N. Vogiatzakis. 2021. "Can Distribution Modeling Inform Rare and Endangered Species Monitoring in Mediterranean Islands?" *Ecological Informatics* 66: 101434.

Naimi, B., N. A. Hamm, T. A. Groen, A. K. Skidmore, and A. G. Toxopeus. 2014. "Where Is Positional Uncertainty a Problem for Species Distribution Modelling?" *Ecography* 37(2): 191–203.

Parmesan, C., and G. Yohe. 2003. "A Globally Coherent Fingerprint of Climate Change Impacts across Natural Systems." *Nature* 421(6918): 37–42.

Pearson, R. G. 2010. "Species' Distribution Modeling for Conservation Educators and Practitioners." *Lessons in Conservation* 3: 54–89.

Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. T. Peterson. 2007. "Predicting Species Distributions from Small Numbers of Occurrence Records: A Test Case Using Cryptic Geckos in Madagascar." *Journal of Biogeography* 34: 102–17.

Pebesma, E. J., and R. S. Bivand. 2005. "Classes and Methods for Spatial Data in R." *R News* 5(2): 9–13. https://cran.r-project.org/doc/Rnews/.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190(3–4): 231–59.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rabinowitz, D. 1981. "Seven Forms of Rarity." In *Biological Aspects of Rare Plant Conservation*, edited by H. Synge, 205–17. New York: Wiley.

Rushton, S. P., S. J. Ormerod, and G. Kerby. 2004. "New Paradigms for Modelling Species Distributions?" *Journal of Applied Ecology* 41(2): 193–200.

Schwinning, S., B. I. Starr, and J. R. Ehleringer. 2005. "Summer and Winter Drought in a Cold Desert Ecosystem (Colorado Plateau) Part I: Effects on Soil Water and Plant Water Uptake." *Journal of Arid Environments* 60(4): 547–66.

Seoane, J., L. M. Carrascal, C. L. Alonso, and D. Palomino. 2005. "Species-Specific Traits Associated to Prediction Errors in Bird Habitat Suitability Modelling." *Ecological Modelling* 185: 299–308.

Silvertown, J., M. Harvey, R. Greenwood, M. Dodd, J. Rosewell, T. Rebelo, J. Ansine, and K. McConway. 2015. "Crowdsourcing the Identification of Organisms: A Case-Study of iSpot." *ZooKeys* 480: 125–46.

Singh, M. 2013. "Predictive Modelling of the Distribution of Two Critically Endangered Dipterocarp Trees: Implications for Conservation of Riparian Forests in Borneo." *Journal of Ecology and The Natural Environment* 5(9): 254–9.

Suzuki-Ohno, Y., J. Yokoyama, T. Nakashizuka, and M. Kawata. 2017. "Utilization of Photographs Taken by Citizens for Estimating Bumblebee Distributions." *Scientific Reports* 7(1): 1–11.

Velazco, S. J. E., F. Galvao, F. Villalobos, and P. De Marco Junior. 2017. "Using Worldwide Edaphic Data to Model Plant Species Niches: An Assessment at a Continental Extent." *PLoS One* 12(10): e0186025.

Walther, B. A., N. Schäffer, A. Van Niekerk, W. Thuiller, C. Rahbek, and S. L. Chown. 2007. "Modelling the Winter Distribution of a Rare and Endangered Migrant, the Aquatic Warbler *Acrocephalus paludicola*." *Ibis* 149(4): 701–14.

Wang, W. C., N. J. Lo, W. I. Chang, and K. Y. Huang. 2012. "Modeling Spatial Distribution of a Rare and Endangered Plant Species (*Brainea insignis*) in Central Taiwan." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39: 241–6.

Wershow, S. T., and E. G. DeChaine. 2018. "Retreat to Refugia: Severe Habitat Contraction Projected for Endemic Alpine Plants of the Olympic Peninsula." *American Journal of Botany* 105(4): 760–78.

Wickham, H., R. François, L. Henry, and K. Müller. 2022. "dplyr: A Grammar of Data Manipulation." R Package Version 1.0.8. https://CRAN.R-project.org/package=dplyr.

Williams, J. N., C. Seo, J. Thorne, J. K. Nelson, S. Erwin, J. M. O'Brien, and M. W. Schwartz. 2009. "Using Species Distribution Models to Predict New Occurrences for Rare Plants." *Diversity and Distributions* 15(4): 565–76.

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*, 2009th ed. New York: Springer.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gaier, Andrew G., and Julian Resasco. 2023. "Does Adding Community Science Observations to Museum Records Improve Distribution Modeling of a Rare Endemic Plant?" *Ecosphere* 14(3): e4419. https://doi.org/10.1002/ecs2.4419