Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth: Don't let the perfect be the enemy of the good

Running title: Using imprecisely geolocated occurrences

Authors:

Adam B. Smith[1*] [ORCiD: 0000-0002-6420-1659]
Stephen J. Murphy[1,2] [ORCiD: 0000-0001-8679-2851]
David Henderson[3] [ORCiD: 0000-0003-3523-0175]
Kelley D. Erickson[1] [ORCiD: 0000-0001-7980-9985]

[1] Center for Conservation and Sustainable Development, Missouri Botanical Garden, 4344 Shaw Boulevard, Saint Louis MO 63110 USA

[2] Present address: The Ohio State University, Department of Evolution, Ecology, and Organismal Biology, 318 W 12th Avenue, Columbus, OH 43210 USA, murphy.1132@osu.edu

[3] Department of Biology, Washington University in Saint Louis, 1 Bookings Drive, Saint Louis MO 63130 USA, hdavid@wustl.edu

* Corresponding author

Abstract [300 words]

Museum and herbarium specimen records are frequently used to assess species' conservation status and responses to climate change. Typically, records of occurrence with imprecise geolocality information are discarded because they cannot be matched confidently to environmental conditions, and are thus expected to increase uncertainty in downstream analyses. However, using only those records that can be precisely georeferenced risks undersampling of species' environmental and geographic distributions. Using simulated and real species, we compared the effect of discarding versus retaining imprecise records on the accuracy of ecological niche models (ENMs), and estimates of niche breadth and extent of occurrence. Imprecise records were assigned to locations or climate values using a conservative approach. For simulated species, including imprecisely-georeferenced records alongside precisely-

georeferenced records improved the accuracy of ENMs projected to the present and the future. Improvements were especially notable for species with <~20 precise occurrences, but accuracy continued to improve and eventually matched the accuracy of ENMs using all records with no georeferencing error. Using only precise records underestimated loss in suitable habitat and overestimated the amount of suitable habitat in both the present and future. Including imprecise records also improved univariate and multivariate estimates of niche breadth, and of species' extent of occurrence. An analysis of 44 species in the genus *Asclepias* (milkweeds) revealed similarly large differences between cases using only precise records and those using precise plus imprecise records. Although the inclusion of imprecisely-georeferenced occurrence records may not always be appropriate, careful consideration must be made regarding the trade-offs between fidelity of the match between occurrences and the environment, versus sample size and how well precisely-geolocated records sample species' environmental and geographic distributions. The preponderance of imprecisely georeferences specimens in museums and herbaria could be gainfully employed to address known shortfalls in sampling of species' distributions and climatic niche tolerances.

Keywords

climate change vulnerability, coordinate uncertainty, georeferencing, niche breadth, natural history museum specimen records, niche truncation, rare species

Introduction

Accurate estimation of species' environmental tolerances and distributions is key to addressing many pressing issues in ecology, evolution, and conservation (e.g., Fisher-Reid et al. 2012; Foden et al. 2013; Quintero & Wiens 2013a and b). Information on species' environmental

tolerances and distributions is commonly inferred from occurrence records obtained from natural history museums and herbaria. These records can be matched to associated environmental conditions and then used to calculate the range of environments inhabited by a species (e.g., Foden et al. 2013). Similarly, ecological niche models (ENMs; also known as species distribution models) can be employed to estimate niche breadth and predict the extent of suitable habitat (Dawson et al. 2011; Pacifici et al. 2015; Foden & Young 2016; Miller et al. 2012; Young et al. 2016; IUCN 2019). Conservation assessments also frequently utilize the extent of occurrence, or area encompassed by occurrence records, to appraise species' potential exposure to broad-scale threats (IUCN 2019; Faber-Langendoen et al. 2012). Methods utilizing opportunistic occurrence records have thus become an important component of biodiversity and conservation research (Heberling 2020). Unfortunately, a large portion of these records is imprecisely geolocated (Moudrý & Devillers 2020), which poses challenges for estimating niches and distributions (Moudrý & Šímová 2012).

For occurrence data to be useful for estimating niche limits and distributions, two key conditions are desirable. First, environmental information assigned to an occurrence record should reflect the conditions that were actually experienced by the species at the location where it was observed (Graham et al. 2008). When records can only be geolocated imprecisely (i.e., to a large region or geopolitical unit), it becomes difficult to confidently assign a specific environmental datum to each record (Feeley & Silman 2010). Second, when the goal is to estimate environmental tolerances, the set of records used to estimate niche limits should encompass as much of the species' fundamental niche as possible (Thuiller et al. 2004; Peterson et al. 2018). If available occurrences represent only a portion of the niche, then the species' actual environmental tolerances will be underestimated (Thuiller et al. 2004; Qiao et al. 2019).

The desire for spatial precision and representative sampling of occurrences leads to a critical trade-off. On the one hand, if records can only be imprecisely geolocated, using them risks introducing uncertainty and bias into estimates of environmental tolerances (Graham et al. 2008; Fernandez et al. 2009; Osborne & Leitão 2009; Feeley & Silman 2010; Tulowiecki et al. 2015; Gábor et al. 2020; Mitchell et al. 2016; Collins et al. 2017; Cheng et al. 2021). On the other hand, discarding records risks under-representing the true geographic and environmental range of a species, even if the location of occurrences is uncertain (Graham et al. 2008). These risks are especially great for rare species, which tend to be represented by just a few records (Lomba et al. 2010; Sheth et al. 2012; Zizka et al. 2018).

To date, these trade-offs have been almost exclusively managed by discarding imprecisely georeferenced records (Moudrý & Šímová 2012). Indeed, in a literature survey of publications using museum or herbarium specimens with ENMs, we found that half of studies that included any description of their data cleaning process discarded spatially imprecise records, and no studies reported purposefully retaining them (Supplementary material Appendix 1 Fig. A2). In other words, the risk of including inaccurate climate data has been assumed more serious than undersampling of the realized niche. However, the justification for discarding spatially imprecise records is often based on studies that do not necessarily reflect real-world use cases. Instead, these studies evaluate the effects of coordinate imprecision on the accuracy of ENMs by artificially adding spatial error to otherwise precise records, and then comparing results between "fuzzed" and accurate records (e.g., Graham et al. 2008; Fernandez et al. 2009; Osborne & Leitão 2009; Gueta & Carmel 2016; Mitchell et al. 2016; Hefley et al. 2017; Soultan & Safi 2017; Tulowiecki et al. 2015; Gábor et al. 2020). This approach, however, is not reflective of the common situation where an assessor starts with a mix of relatively precisely- and imprecisely-

geolocated records and must decide how to delineate the two groups and whether or not to discard the imprecise ones.

Here we reexamine the trade-offs between retaining versus discarding spatially imprecise records using virtual and real species. We designated records as "precise" if they had spatial uncertainty small enough to match them with confidence to environmental data, and "imprecise" if they could not. Imprecision in record coordinates can be represented in a variety of ways, including through the use of user-defined polygons (Wieczorek et al. 2004) or by assignment of records to the smallest geopolitical unit encompassing the area of likely collection (e.g., county, parish, state, etc.; Park & Davis 2017). We emphasize that our definition of an imprecise record does not include records with locations appearing to be outside the range of the species (i.e., geographic outliers; Feeley & Silman 2010) or specimens that do not pass quality-assurance checks (Chapman 2005). We compared ENMs, niche breadth, and the spatial extents of occurrence estimated using only precise records to estimates based on precise plus imprecise records. We hypothesized that the inclusion of imprecise records alongside precise records would improve the accuracy of predicted distributions, estimates of univariate and multivariate niche breadth, and extents of occurrence. We evaluated the accuracy of each of these metrics as a function of how many imprecise records were included, and compared them to benchmark estimates based on "omniscient" records where all populations of a species could be georeferenced without error.

Methods

Study region

For both virtual and real species, the region of analysis encompassed North America (i.e., Canada, the United States, and Mexico, excluding distant islands such as the Hawaiian Islands). However, we calibrated and evaluated ENMs within species-specific regions as noted below.

Virtual species

We used virtual species, for which we could obtain "true" distributions and environmental relationships, to explore the effects of including versus excluding imprecisely geolocated specimens. Full details are found in Fig. 1 and Supplementary material Appendix 2, so are briefly described here. We first calculated a principal component (PC) analysis of all present-day climatic conditions for North America using all 19 BIOCLIM variables (Nix 1986) from WorldClim Version 2.1 at 10 arcmin resolution, which represents average conditions across 1970-2000 (Fick & Hijmans 2017). Species' niches were generated using a trivariate Gaussian distribution in the first three PC axes, then projected to North America (Fig. 1, steps). We interpreted these values as probabilities of presence.

We then created occurrence data for each species by sampling raster cells according to their probability of occurrence times cell area. We first drew a set of "omniscient" records, which reflect an ideal case where a collector is able to georeference every single population of a species with negligible error. We generated species with 20, 40, 80, 160, and 320 omniscient records. From the omniscient records we randomly sampled a subset of 5, 10, 15, 20, 25, or 30 to represent "precise" records (so long as the number of precise records was less than the number of omniscient records). Precise records represent a realistic case where a collector has access to a subset of the species' occurrences georeferenced with little error. To represent "imprecise" records that could only be georeferenced to a county, parish, or equivalent, from the remainder of

omniscient records we sampled a subset of 1, 2, 4, 8, …, records, up to the total number of omniscient records minus the number of precise records. We then assigned these records to the county in which they fell.

For each combination of number of omniscient, precise, and imprecise records, we generated 200 species. We then constructed ENMs and estimated climatically suitable area and niche breadth for each species in three separate ways. First, we calibrated an ENM and niche breadth metrics using all omniscient occurrences to represent a baseline against which to assess the other two sets of models and metrics. Second, we calculated a "precise-only" ENM and metrics using just the precise records, recreating the common practice of discarding imprecise records before biogeographic analysis. Third, we combined the precise and imprecise records to generate "precise + imprecise" ENMs and metrics.

ENMs: We modeled each species' realized niche using MaxEnt (Phillips et al. 2006; Phillips & Dudík 2008) with linear, quadratic, and interaction terms, with the optimal set of terms and master regularization parameter identified using AIC_c (Warren & Siefert 2011). Models were projected to the present and to a future climate scenario defined by the average across five earth system models for 2061-2080 from CMIP6 under Representative Concentration Pathway 8.5 (achievable under Shared Socioeconomic Pathway 5-8.5; O'Neill et al. 2015). We assessed calibration accuracy (the degree to which model output reflects the true probability of presence) of the model predictions in the present and the future using the Pearson correlation coefficient between the ENM predictions and the actual probability of presence. Background sites for model calibration and assessment were drawn from a 300-km buffer around the minimum convex polygon constructed around all omniscient points. We also assessed area of climatically suitable habitat within this same region by applying a threshold such that training sensitivity was 0.9. We

calculated the area of current and future favorable habitat, and the area of habitat that was lost, gained, or remained favorable. To determine how inclusion of imprecise records affected model complexity, we calculated the number of non-zero coefficients (ignoring the intercept) for each type of model.

Niche breadth and extent of occurrence: We also calculated a set of metrics broadly used in ecology, evolution, and conservation: (a) univariate niche breadth in mean annual temperature (MAT) and total annual precipitation (TAP), estimated from the range of the inner 90% of values of each variable across each set of occurrences; (b) multivariate niche breadth, represented by the volume and surface area of a minimum convex hull around each set of occurrences in three-dimensional environmental space; and (c) extent of occurrence (EOO), which is the area of the minimum convex polygon encompassing all records in geographic space (IUCN 2019). EOO is used in conservation assessments as an index of the propensity for multiple populations to experience the same broad scale threat (IUCN 2019).

Assigning imprecise records locations and environments: Since the imprecise records could not, by design, be geolocated with confidence to a single cell, we used a simple yet conservative approach to assign climatic conditions to these records. For the ENMs and multivariate niche breadth, we used the PC score of the cell in each imprecise record's county that was closest in PC space to the mean of the precise records' PC scores. To calculate univariate niche breadth, we used the value of MAT or TAP from across each imprecise record's county that was closest to the mean across the precise records (Fig. 2b and c). To calculate extent of occurrence, we assigned each imprecise occurrence to the point in the county where it was located closest to the centroid of the precise records (Fig. 2a). For example, the actual location of an imprecise record

in a county was almost always further from the precise records' centroid than the closest point in the county.

We did not formally compare results between data types (omniscient, precise, precise plus imprecise) using statistical tests since these are inappropriate for simulations where the existence of differences is known a priori (White et al. 2014). Rather, we compared the inner 90^{th}-percentile distribution of each metric (calibration accuracy, EOO, etc.) for each case.

Real species

We also evaluated 44 species of *Asclepias* (milkweeds; family Apocynaceae) native to North America, which display a range distributions from narrowly endemic to those covering approximately one-third of the continent. Records were obtained from the Global Biodiversity Information Facility (www.gbif.org), and detailed procedures for data cleaning and modeling are described in Supplementary material Appendices 3 and 4. For this analysis we used only herbarium specimens and species with ≥5 precise and unique (i.e., non-duplicate) records. We evaluated the same set of metrics as for the virtual species and created ENMs following the same procedures. However, because we did not have omniscient records for the real species, we estimated climatically suitable habitat area within the region defined by a 300-km buffer around the minimum convex polygon surrounding all available records. We tested for differences in EOO and in univariate and multivariate niche breadth calculated with or without imprecise records using a paired Wilcoxon signed-rank test.

Reproducibility

The analysis relied primarily on the sp (Bivand et al. 2013), rgeos (Bivand & Rundel 2020), geosphere (Hijmans 2019), dismo (Hijmans et al. 2017), raster (Hijmans 2021), and enmSdm (Smith 2021) packages for R Version 4.10 (R Core Team 2021). Scripts used in this analysis are available on the GitHub repository https://github.com/adamlilith/enms_impreciseRecords.

Results

Virtual species

For brevity, we focus on results for species with 40 and 320 omniscient records, each with 5 or 20 precise records (see Supplementary material Appendix 2 for all results). We found that adding imprecise records caused the greatest improvements in metrics when the number of precise records was <15-20, although the accuracy of some measures continued to improve as imprecise records were added. To reflect real-world use cases where assessors may have a set of records that can only be located to a geopolitical unit but not otherwise know if they represent truly geographically unique specimens, we plotted the change in each metric (ENM accuracy, niche breadth, etc.) against the number of county-level records accrued as imprecise records were added. Hence, hereafter we refer to changes in metrics as a function of the number of counties encompassing the imprecise records (counties are only counted once, regardless of the number of imprecise records they contain).

ENMs: Compared to precise-only models, precise + imprecise models were more accurate when projected to both present and future climate scenarios (Figs. 3a and b, and Appendix 2 Figs. A1 and A2). Accuracy increased most rapidly when the number of precise records in precise + imprecise models was ≤15, but continued to rise thereafter. For example, for present-day climate and a species with 40 total occurrences and only 5 precise occurrences, the median correlation

224 between the true probability of presence and predictions from precise-only ENMs was 0.15 (90%
225 inner quantiles: -0.02 to 0.58). Adding county records increased the correlation up to 0.63 (0.30-
226 0.86), which was equivalent to the accuracy of omniscient ENMs (median: 0.60, 90% quantile:
227 0.27-0.81). Similar results were obtained for species with a greater number of total occurrences.
228 When the number of precise records was ≥20 or when >~30 county records were included,
229 increases in accuracy of precise + imprecise models were more gradual (Fig. 3a). However, for
230 species with ≥160 total occurrences, there was a second notable increase in accuracy when the
231 number of county records surpassed about half of the total number of omniscient records (Fig. 3a
232 and b). Accuracy decreased when ENMs were projected to the future, although precise +
233 imprecise models still outperformed precise-only models when a sufficient number of county
234 records was included (Fig. 3b). Precise-only models were simpler (fewer non-zero coefficients)
235 than omniscient and precise + imprecise models (Supplementary material Appendix 2 Fig. A13).

236 Compared to estimates based on omniscient records, precise-only ENMs overestimated the area
237 of climatically suitable habitat by up to >300% (median value) for the present (Fig. 3c) and by
238 >500% for the future (Fig. 3d and Supplementary material Appendix Figs. A6 and A7).
239 Overestimation was greatest when the number of precise occurrences was <15. Adding just ~10-
240 30 county-level records reduced bias to ~0 relative to omniscient models.

241 When the number of precise records was ≤15, precise-only ENMs underestimated loss of
242 suitable area by 90% or more (median; Fig. 3e and f, Supplementary material Appendix 2 Fig.
243 A3). Adding 10-30 county records eliminated this bias. Gains in suitable area were on average
244 unbiased (Supplementary material Appendix 2 Fig. A4). Precise-only ENMs overestimated area
245 that remained climatically suitable through time (Supplementary material Appendix 2 Fig. A5).

246 *Niche breadth*: Using just precise records underestimated univariate niche breadth and
247 multivariate niche volume and surface area (Fig. 4b-e and Supplementary material Appendix 2
248 Figs. A8 to A11). For example, for a species with 40 total occurrences of which just 5 could be
249 geolocated precisely, precise-only niche breadth in MAT was only 29% (6-71%) of omniscient
250 niche breadth. However, precise + imprecise niche breadth in temperature increased up to ~90%
251 (~70-110%) of omniscient breadth as county records were added. For species with ≥20 precise
252 occurrences, improvements in estimates of niche breadth were less dramatic. Similar trends were
253 observed in niche breadth in TAP, and in multivariate niche volume and surface area (Fig. 4c-d).

254 Occasionally, adding imprecise data led to overestimation, especially for niche breadth in MAT
255 and niche volume. For MAT, across all species and scenarios, 4.5% of precise + imprecise
256 estimates were larger than omniscient values, with the median of overestimated values 5% larger
257 than the true value. For niche volume, 9% of precise + imprecise estimates were larger than
258 omniscient values, with the median of overestimated values 27% larger than the true value. For
259 MAT, overestimation was more common when the number of precise records was large and
260 additional county records was small (Fig. 4b), but for niche volume overestimation was more
261 likely as county records increased (Fig. 4d).

262 *Extent of occurrence*: EOO was consistently underestimated when only precise records were
263 used (Fig. 4a and Supplementary material Appendix 2 Fig. A12). Underestimation was worse for
264 species with few precise records and a large number of occurrences. In these cases, median
265 estimated EOO was as small as 12% of the actual value (inner 90% quantile range: 2-37%).
266 Adding county records improved estimates in nearly every case, though the rate of improvement
267 declined as more county records were added. EOO was occasionally overestimated when county

268 records were included. Across all scenarios, 1.7% of precise/imprecise estimates were larger than
269 omniscient values, with the median of overestimate values 47% larger than the true value.

270 *Asclepias*

271 The data obtained from GBIF comprised 53,623 herbarium records. Following data cleaning,
272 removal of observational and duplicate records, and elimination of species with fewer than 5
273 geographically unique precise records, we were left with just 16% of the original records (8,480)
274 and only 32% of the species (44 of 137). Imprecise records were the most abundant, comprising
275 on average 70% of all usable records for a species (range: 44-97%; Supplementary material
276 Appendix 3 Table A1). Thirty-six percent (16 of 44) of species had ≤20 precise records and
277 twenty percent (9 of 44) had ≤10.

278 *ENMs*: Relative to precise-only ENMs, precise + imprecise models predicted greater climatically
279 suitable area for 63% of species (28 of 44) in the present and 71% of species (31 of 44) in the
280 future (Fig. 5a and b, Supplementary material Appendix A7 Figs. A3). Gains and losses in
281 suitable area were also larger when imprecise records were included (median gain 16% ranging
282 from -80 to 3952%; median loss 30% ranging from -89 to 1652%; Supplementary material
283 Appendix A7 Fig. A5).

284 *Niche breadth*: Including imprecise records increased median univariate niche breadth in MAT
285 and TAP by 25% across species (range: 0 to 353%) and 28% (0 to 292%), respectively (Fig. 5d;
286 Wilcoxon V ~0 and $P<10^{-6}$ in both cases). Including imprecise records increased multivariate
287 niche volume by a median value of 175% (8 to 13,909%; $P<10^{-12}$, Wilcoxon V~0) and niche
288 surface area by 79% (3 to 1515%; $P<10^{-12}$, Wilcoxon V~0; Fig. 5e). Species with the fewest

289 records had the greatest increase in niche breadth, volume, and area when imprecise records were
290 included.

291 *Extent of occurrence*: Precise + imprecise EOO was 86% larger (median; range 0-2011%) than
292 precise-only EOO (Fig. 5c). Including imprecise records at least doubled EOO for 34% of
293 species (15 of 44), and at least tripled EOO for 27% of species (12 of 44). Species with fewer
294 precise records tended head greater increases in EOO when imprecise records were included.

295 **Discussion**

296 Spatially imprecise occurrences are commonly discarded prior to biogeographic analyses
297 (Appendix 1 Fig. S2). Although the justification for this decision is due to the potential for
298 spatial error to propagate through an analysis (Moudrý & Šímová 2012), discarding records also
299 risks introducing error by undersampling geographic and environmental space. We found that
300 including geospatially-imprecise records increased the accuracy of niche models projected to
301 both present and future climate conditions (Fig. 3a and b). Including these records also led to
302 more accurately estimated climatically suitable area (Fig. 3c-e) and improved estimates of niche
303 breadth and extent of occurrence (EOO; Fig. 4). Sometimes estimates using just precise records
304 were orders-of-magnitude different from those based on "omniscient" records, but adding
305 imprecise records helped close this gap.

306 A critical distinction of our work is that we show the effects of adding imprecise records to
307 precise records, whereas most studies on coordinate uncertainty demonstrate the effects of
308 making precise records spatially imprecise (e.g., Graham et al. 2008; Fernandez et al. 2009;
309 Soultan & Safi 2017; Gábor et al. 2020). Likewise, studies that have examined the effects of
310 adding additional imprecise records to ENMs have typically not employed a conservative

method for assigning environmental values or locations to imprecise records (e.g., Bloom et al. 2018; Collins et al. 2017; Cheng et al. 2021), or have focused on mean values (versus extremes, which define environmental limits; Pender et al. 2019). Our findings have direct implications for studies in ecology, evolution, and conservation that use occurrences to estimate species' relationships to the environment. More broadly, the uncertainty inherent in the decision over whether to use or discard imprecise records should be reflected in the outcome of analyses relying on occurrence data.

Why including imprecise records improves niche models and estimates of niche breadth

The effect of imprecise records on ENM accuracy and niche breadth depends on the abundance of precise records and how well they sample geographic and niche space (Moudrý & Šímová 2012; Tulowiecki et al. 2015; Soultan & Safi 2017). We found that including a sufficient number of imprecise records could fully compensate for a lack of precise records even when there were as few as 5 precise occurrences. Gains in accuracy were most notable for species with <~20 precise records, but even species >20 precise occurrences experienced improvements from adding imprecise records. Sample size is one of the largest influences on niche model accuracy (Santini et al. 2021), with minimum recommended sizes ranging from <10 to several hundred (Wisz et al. 2008; van Proossdij et al. 2016; Santini et al. 2021; see also Rivers et al. 2011). Given that many species are known from just a handful of records (Zizka et al. 2018), including imprecise records could be especially helpful when sample sizes are small. Adding imprecise records can improve the sampling of geographic and environmental space (e.g., Fig. 2a).

Implications for studies in evolution, ecology, and conservation

Although the amount of biodiversity data is growing at an astounding rate, most species' distributions remain poorly characterized (Meyer et al. 2016). For North American *Asclepias*, even with >53,000 specimen records we had enough data to analyze only a third species found in North America. Of those we did analyze, 70% of their records would have normally been discarded due to spatial uncertainty in their locations. Similar rates of spatial uncertainty are common in biodiversity databases (Moudrý & Devillers 2020).

Spatially imprecise records, if used carefully, have great potential to address the "Wallacean shortfall," the lack of information on species' distributions (Hortal et al. 2015), and the "Hutchinsonian" shortfall, the lack of information on species' environmental tolerances (Cosentino & Maiorano 2021). Answers to many key questions in ecology and evolution are susceptible to these shortfalls, and so could be informed by addition of erstwhile "unusable" imprecise records. For example, inadequate representation of environmental conditions inhabitable by a species can bias estimates of the rate of climatic niche evolution (Saupe et al. 2018) and alter species' response curves along environmental gradients (Hannemann et al. 2016). Likewise, investigations of relationships between niche breadth, geographic range size and environmental variation (Quintero & Wiens 2013a and b), and measurements of niche overlap (Warren et al. 2008) will be inherently sensitive to the degree to which realized niches are adequately sampled. The many studies that rely on ENMs for reconstructing species' past, present, and potential future distributions are especially sensitive to sample size (Santini et al. 2021) and uneven sampling intensity among inhabitable environments (Raes 2012; Perret & Sax 2022).

Discarding imprecise records has the potential to overestimate species' vulnerability and thus bias conservation assessments. For example, under IUCN Red List criterion B1, species qualify

as threatened if they have an EOO <100 km^2 (in addition to other criteria; IUCN 2019; see also Young et al. 2016). While none of the 44 species of *Asclepias* in our analysis had an EOO <100 km^2 when using just precise records, it is certainly possible that this threshold could be crossed by some of the other 93 species in our original data that we did not analyze because they had <5 geographically unique precise records. Similarly, estimates of species' adaptive capacities (Cang et al. 2016), rates of community thermophilization (Feeley et al. 2020), and exposure to anticipated climate change (Fig. 3) could be misrepresented by undersampling due to removal of imprecise specimens.

Our intent is to provoke a reconsideration of the benefits and costs of discarding spatially imprecise records when assessing a species' response to environmental gradients and geographic distributions. These trade-offs must be assessed within the goals and philosophical approach of an analysis. For example, many conservation assessments adopt a precautionary strategy that errs on the side of assuming a species is more vulnerable than it may actually be (Moyle 2005; Huntley et al. 2016a; IUCN 2019). In contrast, an evidentiary approach aims to classify species as vulnerable only if there is strong evidence to support such a designation (IUCN 2019). Discarding imprecise records decreased estimates of niche breadth and EOO (Fig. 4), so aligns with a precautionary approach because the species appear more vulnerable then they may be. However, using just precise records did overestimate climatically suitable area for the virtual species (Fig. 3c and 3d). In contrast, niche breadth and EOO were occasionally overestimated when imprecise records were included (Fig. 4). Hence, whether or not a decision to retain versus keep imprecise records is precautionary or evidentiary depends on the metric used to assess vulnerability.

In this context, we did find some differences between virtual and real species. Specifically, including imprecise records reduced the estimated climatically suitable area in the present and future for virtual species, but increased it for real species (Fig. 3c and d versus Fig. 5d and f; cf. Collins et al. 2017). Inspection of the individual ENMs for virtual species revealed that models that overestimated suitable habitat were: a) largely relegated to cases with small sample size (total number of precise or precise plus imprecise records) and small EOO; and b) were much simpler (i.e., had only a single term or were intercept-only models; Supplementary material Appendix 2 Fig. A13). Small sample size favors simpler models because available information does not justify complex responses (Phillips et al. 2006; Warren & Siefert 2011; Merow et al. 2014; Vignali et al 2020). However, simple models containing, say, just linear terms cannot represent unimodal responses along environmental gradients (Whittaker 1953). Rather, they predict that suitability increases along a gradient, even if conditions eventually become worse again. As a result, and somewhat ironically, overly simple models tended to predict that suitable habitat for species with few occurrences is larger than for the same species with more occurrences, which allowed for more complex models (Brun et al. 2020). We did not see this effect in the real species, but 70% of our real species (31 of 44) had >15 precise records, so models were less likely to be simple.

Uncertainty

Analyses are most informative when they account for all relevant aspects of uncertainty (Huntley et al. 2016b; IUCN 2019). The decisions over whether to retain or discard imprecise records, and indeed, over what constitutes an "imprecise" record, represent key aspects of uncertainty. However, across all of the articles that we reviewed, none of them evaluated the consequences of discarding imprecise records (Supplementary material Appendix 1 Fig. A2). Ignoring the

subjectivity inherent in these decisions reduces apparent uncertainty in the final assessment of species' distributions and environmental tolerances. Acknowledging uncertainty is important even in cases where the exclusion of imprecise records would not change a species' overall vulnerability status, since increased uncertainty reduces the reliability of returns on conservation investment (Smith et al. 2016).

Conclusions

We advocate for a re-consideration over whether spatially imprecise occurrence records should be excluded from biogeographic analyses. Using only precise records reduces niche model accuracy, and can underestimate niche breadth and extent of occurrence. The decision over how to define imprecise records and whether or not to use them is an important contributor to the overall uncertainty in inherent in the outcome of analyses relying on specimen data. Discarding imprecise records ignores a critical aspect of uncertainty and risks undersampling species' realized environmental and geographic distributions. Practitioners need to consider the trade-offs between using versus discarding imprecise records, especially given the preponderance of imprecise records available in specimen databases and the Wallacean and Hutchinsonian shortfalls that beset our sampling of the distribution of life on Earth.

Acknowledgements

DH was supported by the Department of Biology of Washington University in Saint Louis. SJM and KDE were supported by the Institute of Museum and Library Services (FAIN MG-30-15-0094-15) to ABS. This research was also supported by the Alan Graham Fund in Global Change.

Conflict of Interest Statement

421 The authors declare no conflict of interest.

422 **Biosketch**

423 The Global Change Conservation Laboratory at the Missouri Botanical Garden identifies
424 solutions to pressing environmental problems by leveraging precise and the imprecise
425 information on biodiversity. We are inspired by the cumulative person-millennia of field work
426 and curatorial attention devoted to amassing collection records of Earth's species and the
427 exigency of using this data to its fullest potential.

428 **Literature cited**

429 Bivand, R. and Rundel, C. 2020. rgeos: Interface to Geometry Engine: Open Source ('GEOS').
430 Version 0.5-5. https://CRAN.R-project.org/package=rgeos
431 Bivand, R., Pebesma, E., and Gómez-Rubio, V. 2013. Applied Spatial Data Analysis with R.
432 Springer-Verlag, New York.
433 Bloom, T.D.S., Flower, A., and DeChaine, E.G. 2018. Why georeferencing matters:
434 Introducing a practical protocol to prepare species occurrence records for spatial analysis.
435 Ecology and Evolution 8:765-777.
436 Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R.O., Wang, Z., and Zimmermann,
437 N.E. 2020. Model complexity affects species distribution projections under climate change.
438 Journal of Biogeography 47:130-142.
439 Cang, F.A., Wilson, A.A., and Wiens, J.J. 2016. Climate change is projected to outpace rates of
440 niche change in grasses. Biology Letters 12: 20160368.
441 Chapman, A.D. 2005. Principles and Methods of Data Cleaning: Primary Species and Species-
442 Occurrence Data, Version 1.0. Report for the Global Biodiversity Information Facility,
443 Copenhagen.
444 Cheng, Y., Tjaden, N.B., Jaeschke, A., Thomas, S.M, and Beierkuhnlein, C. 2021. Using
445 centroids of spatial units in ecological niche modeling: Effects on model performance in the
446 context of environmental data grain size. Global Ecology and Biogeography 30:611-621.

447 Collins, S.D., Abbott, J.C., and McIntyre, N.E. 2017. Quantifying the degree of bias from using
448 county-scale data in species distribution modeling: Can increasing sample size or using
449 county-averaged environmental data reduce distributional overprediction? Ecology and
450 Evolution 7:6012-6022.
451 Cosentino, F. and Laiorano, L. 2021. Is geographic sampling bias representative in
452 environmental space? Ecological Informatics 64:101369.
453 Dawson, T.P., Jackson, S.T., House, J.I., Prentice, I.C., and Mace, G.M. 2011. Beyond
454 predictions: Biodiversity conservation in a changing climate. Science 332:53-58.
455 Faber-Langendoen, D., J. Nichols, L. Master, K. Snow, A. Tomaino, R. Bittman, G. Hammerson,
456 B. Heidel, L. Ramsay, A. Teucher, and B. Young. 2012. NatureServe Conservation Status
457 Assessments: Methodology for Assigning Ranks. NatureServe, Arlington, VA.
458 Feeley, K.J. and Silman, M.R. 2010. Modelling the responses of Andean and Amazonian plant
459 species to climate change: The effects of georeferencing errors and the importance of data
460 filtering. Journal of Biogeography 37:733-740.
461 Feeley, K.J., Bravo-Avila, C., Fadrique, B., Perez, T.M., and Zuleta, D. 2020. Climate-driven
462 changes in the composition of New World plant communities. Nature Climate Change
463 10:965-970.
464 Fernandez, M.A., S.D. Blum, S. Reichle, Q. Guo, B. Holzman, and H. Hamilton. 2009. Locality
465 uncertainty and the differential performance of four common niche-based modeling
466 techniques. Biodiversity Informatics 6:36-52.
467 Fick, S.E. and Hijmans, R.J. 2017. WorldClim 2: New 1-km spatial resolution climate surfaces
468 for global land areas. International Journal of Climatology 37:4302-4315.
469 Fisher-Reid, M.C., Kozak, K.H., and Wiens, J.J. 2012. How is the rate of climatic-niche
470 evolution related to climatic-niche breadth? Evolution 66:3836-3851.
471 Foden, W.B. and Young, B.E. (eds.) 2016. IUCN SSC Guidelines for Assessing Species'
472 Vulnerability to Climate Change. Version 1.0. Occasional Paper of the IUCN Species
473 Survival Commission No. 59. Cambridge, UK and Gland, Switzerland: IUCN Species
474 Survival Commission. x+114 pp.
475 Foden, W.B., Mace, G.M., and Butchart, S.H.M. 2013. Indicators of climate change impacts on
476 biodiversity. Pp. 120-137 in Collin, B., Pettorelli, N., Baillie, J.E.M., and Durant, S.M.

(eds.) Biodiversity Monitoring and Conservation: Bridging the Gap between Global Commitment and Local Action, 1st ed. John Wiley and Sons, Indianapolis.

Gábor, L., Moudrý, V., Lecours, V., Malavasi, M., Barták, V., Fogl, M., Šímová, P., Rocchini, D., and Václavík, T. 2020. The effect of positional error on fine scale species distribution models increases for specialist species. Ecography 43:256-269.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B.A., and the NCEAS Predicting Species Distributions Working Group. 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology 45:239-247.

Gueta, T. and Carmel, Y. 2016. Quantifying the value of user-level cleaning for big data: A case study using mammal distribution models. Ecological Informatics 34:139-145.

Hannemann, H., Willis, K.J., and Macias-Fauria, M. 2016. The devil is in the details: Unstable response functions in species distribution models challenge bulk ensemble modeling. Global Ecology and Biogeography 25:26-35.

Heberling, J.M. 2020. Global change biology: Museum specimens are more than meet the eye. Current Biology 30: R1368-R1370.

Hefley, T.J., Brost, B.M., and Hooten, M.B. 2017. Bias correction of bounded location errors in presence-only data. Methods in Ecology and Evolution 8:1566-1573.

Hijmans, R.J. 2016. raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster.

Hijmans, R.J. 2019. geosphere: Spherical Trigonometry. R package version 1.5-10. https://CRAN.R-project.org/package=geosphere.

Hijmans, R.J., Phillips, S.J., Leathwick, J., and Elith, J. 2020. dismo: Species Distribution Modeling. R package version 1.3-3. https://CRAN.R-project.org/package=dismo.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinson, T.M., Lobo, J.M., and Ladle, R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics 46:523-549.

Huntley, B., Foden, W.B., Pearce-Higgins, J., and Smith, A.B. 2016. Chapter 6. Understanding and working with uncertainty. In W.B. Foden and B.E. Young, editors. IUCN SSC Guidelines for Assessing Species' Vulnerability to Climate Change. Version 1.0. Occasional

Paper of the IUCN Species Survival Commission No. 59. Gland, Switzerland and Cambridge, UK. pp 49-56.

Huntley, B., Foden, W.B., Smith, A.B., Platts, P., Watson, J. and Garcia, R.A. 2016. Chapter 5. Using CCVAs and interpreting their results. In W.B. Foden and B.E. Young, editors. IUCN SSC Guidelines for Assessing Species' Vulnerability to Climate Change. Version 1.0. Occasional Paper of the IUCN Species Survival Commission No. 59. Gland, Switzerland and Cambridge, UK. pp 33-48.

IUCN Standards and Petitions Committee. 2019. Guidelines for Using the IUCN Red List Categories and Criteria. Version 14. Prepared by the Standards and Petitions Committee. Downloadable from http://www.iucnredlist.org/documents/RedListGuidelines.pdf (2021-05-24).

Lomba, A., L. Pellissier, C. Randin, J. Vicente, J. Horondo, and A. Guisan. 2010. Overcoming the rare species modeling complex: A novel hierarchical framework applied to an Iberian endemic plant. Biological Conservation 143:2647-2657.

Merow, C., Allen, J.M., Aiello-Lammens, M., and Silander, Jr., J.A. 2016. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. Global Ecography and Biogeography 25:1022-1036.

Merow, C., Smith, M.J., Edwards, Jr., T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E., and Elith, J. 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37:1267-1281.

Miller, J.S., Porter-Morgan, H.A., Stevens, H., Boom, B., Krupnick, G.A., Acevedo-Rodríguez, P., Fleming, J., and Gensler, M. 2012. Addressing targets two of the Global Strategy for Plant Conservation by rapidly identifying plants at risk. Biodiversity Conservation 21:1877-1887.

Mitchell, P.J., Monk, J., and Laurenson, L. 2016. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across sample sizes. Methods in Ecology and Evolution 8:12-21.

Moudrý, V. and Devillers, R. 2020. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecological Informatics 56:101051.

537 Moudrý, V. and Šímová, P. 2012. Influence of positional accuracy, sample size and scale on
538 modeling species distributions: A review. International Journal of Geographic Information
539 Science 26:2083-2095.
540 Moyle, B. 2005. Making the Precautionary Principle work for biodiversity: Avoiding perverse
541 outcomes in decision-making under uncertainty. Pp. 159-172 in Cooney, R. and B. Dickson
542 (eds.) Biodiversity and the Precautionary Principle: Risk and Uncertainty in Conservation
543 and Sustainable Use. Earthscan, London. 314 pp.
544 Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. Atlas of elapid snakes of
545 Australia: Australian flora and fauna series 7 (ed. by R. Longmore), pp. 4-15. Bureau of
546 Flora and Fauna, Canberra.
547 O'Neill, B.C., Kreigler, E., Ebi, K.L., Kemp-Benedict, E., Riahi, K., Rothman, D.S., van
548 Ruijven, B., van Vuuren, D.P., Birkman, J., Kok, K., Levy, M., and Solecki, W. 2017. The
549 roads ahead: Narratives for shared socioeconomic pathways describing world futures in the
550 21st century. Global Environmental Change 42:169-180.
551 Osborne, P.E. and Leitão, P.J. 2009. Effects of species and habitat positional errors on the
552 performance and interpretation of species distribution models. Diversity and Distributions
553 15:671-681.
554 Pacifici, M., Foden, W.B., Visconti, P., Watson, J.E.M., Butchart, S.H.M., Kovacs, K.M.,
555 Scheffers, B.R., Hole, D.G., Martin, T.G., Akçakaya, H.R., Corlett, R.T., Huntley, B.,
556 Brickford, D., Carr, J.A., Hoffmann, A.A., Midgley, G.F., Pearce-Kelly, P. Pearson, R.G.,
557 Williams, S.E., Willis, S.G., Yoing, B., and Rondinini, C. 2015. Assessing species
558 vulnerability to climate change. Nature Climate Change 5:215-225.
559 Park, D.S. and Davis, C.C. 2017. Implications and alternatives of assigning climate data to
560 geographical centroids. Journal of Biogeography 44:2188-2198.
561 Pender, J.E., Hipp, A.L., Hahn, M., Kartesz, J., Nishino, M., and Starr, J.R. 2019. How
562 sensitive are climatic niche inferences to distribution data sampling? A comparison of Biota
563 of North America Program (BONAP) and Global Biodiversity Information Facility (GBIF)
564 datasets. Ecological Informatics 54:100991.
565 Perret, D.L. and Sax, D.F. 2022. Evaluating a study design for optimal sampling of species'
566 climatic niches. Ecography 2022:e06014.

Peterson, A.T., Cobos, M.E., and Jiménez-García, D. 2018. Major challenges for correlational ecological niche model projections to future climate conditions. Annals of the New York Academy of Sciences 1429:66-77.

Phillips, S.J. and Dudík, M. 2008. Modeling species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography 31:161-175.

Phillips, S.J., Anderson, R.P., and Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231-259.

Qiao, H., Feng, X., Escobar, L.E., Peterson, A.T., Soberón, J., Zhu, G., and Papeş. 2019. An evaluation of transferability of ecological niche models. Ecography 42:521-534.

Quintero, I. and Wiens, J.J. 2013a. What determines the climatic niche width of a species? The role of spatial and temporal climatic variation in three vertebrate clades. Global Ecology and Biogeography 22:422-432.

Quintero, I. and Wiens, J.J. 2013b. Rates of projected climate change dramatically exceed past rates of climate niche evolution among vertebrate species. Ecology Letters 16:1095-1103.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raes, N. 2012. Partial versus full species distribution models. Natureza and Conservação 10:127-138.

Rivers, M.C., Taylor, L., Brummitt, N.A., Meagher, T.R., Roberts, D.L., and Lughadha, E.N. 2011. How many herbarium specimens are needed to detect threatened species? Biological Conservation 144:2541-2547.

Santini, L., Benítez-López, Maiorano L., Čengič, M., and Huijbreghts, M.A. 2021. Assessing the reliability of species distribution projections in climate change research. Diversity and Distributions 27:1035-1050.

Saupe, E.E. Barve, N., Owens, H.L., Cooper, J.C., Hosner, P.A., and Peterson, A.T. 2018. Reconstructing niche evolution when niches are incompletely characterized. Systematic Biology 67:428-438.

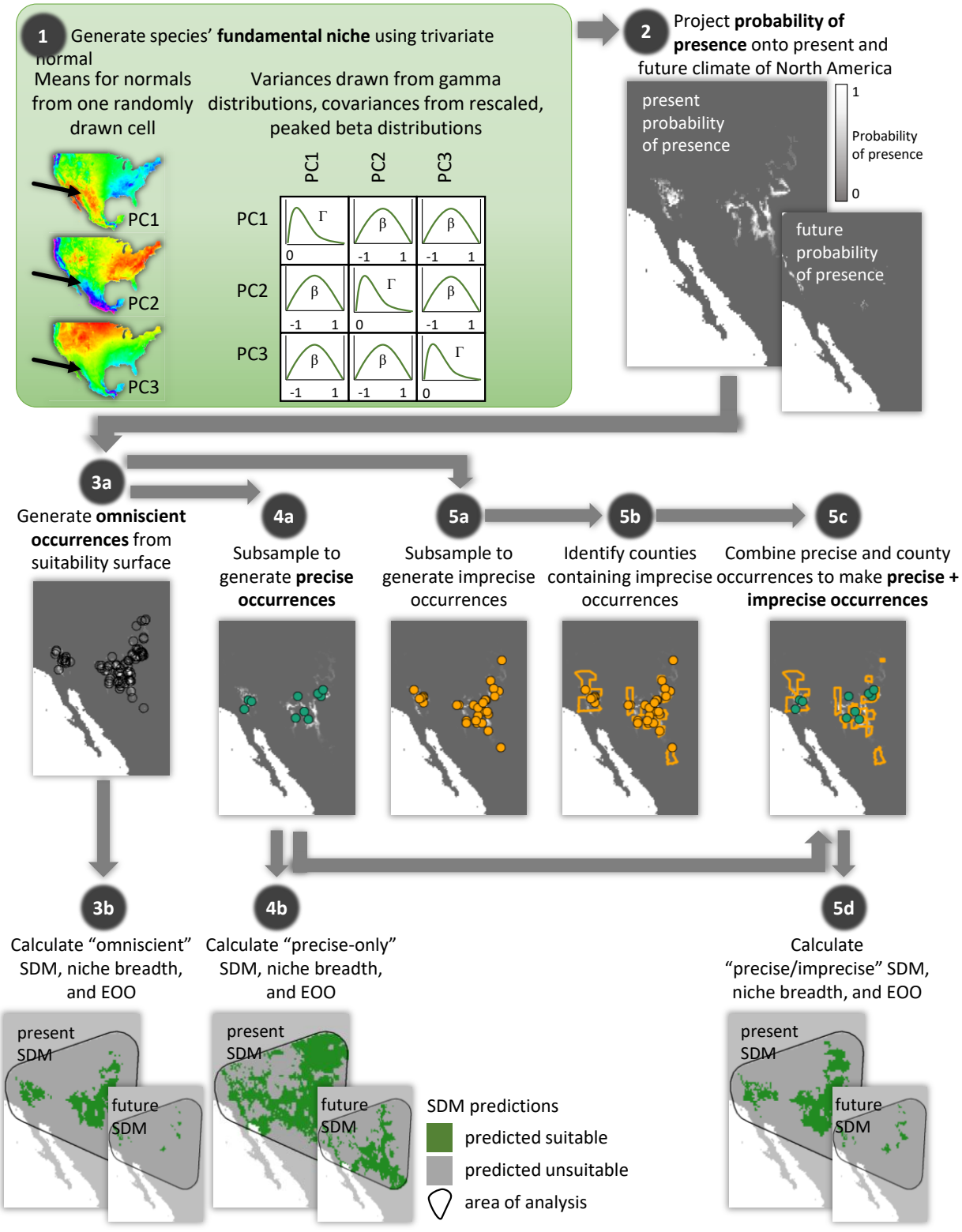
Sheth, S.N., L.G. Lohmann, T. Distler, and I. Jiménez. 2012. Understanding bias in geographic range size estimates. Global Ecology and Biogeography 21:732-742.

Smith, A.B. 2021. enmSdm: Tools for modeling niches and distributions of species. R package version 0.7.0. http://github.com/adamlilith/enmSdm

598 Smith, A.B., Long, Q.G., and Albrecht, M.A. 2016. Shifting targets: spatial priorities for ex situ
599 plant conservation depend on interactions between current threats, climate change, and
600 uncertainty. Biodiversity and Conservation 25:905-922.
601 Soultan, A. and Safi, K. 2017. The interplay of various sources of noise and reliability on
602 species distribution models hinges on ecological specialization. Public Library of Science
603 ONE 12:e0187906.
604 Thuiller, W., Araújo, M.B., Pearson, R.G., Whittaker, R.J., Brotons, L., and Lavorel, S. 2004.
605 Uncertainty in predictions of extinction risk. Nature 430:33.
606 Tulowiecki, S.J., Larsen, C.P.S., and Wang, Y-C. 2015. Effects of positional error on modeling
607 species distributions: A perspective using presettlement land survey records. Plant Ecology
608 216:67-85.
609 van Proossdij, A.S.J., Sosef, M.S.M., Wieringa, J.J. and Raes, N. 2016. Minimum number of
610 specimen records to develop accurate species distribution models. Ecography 39:542-552.
611 Vignali, S., Barras, A.G., Arlettaz, R. and Braunisch, V. 2020. SDMtune: An R package to tune
612 and evaluate species distribution models. Ecology and Evolution 10:11488-11506.
613 Warren, D.L. and S.N. Siefert. 2011. Ecological niche modeling in Maxent: The importance of
614 model complexity and the performance of model selection criteria. Ecological Applications
615 21:335-342.
616 Warren, D.L., Glor, R.E., and Turelli, M. 2008. Environmental niche equivalency versus
617 conservatism: Quantitative approaches to niche evolution. Evolution 62:2868-2883.
618 White, J.W., Rassweiler, A., Samhouri, J.F., Stier, A.C., and White, C. 2014. Ecologists should
619 not use statistical significance tests to interpret simulation model results. Oikos 123:385-388.
620 Whittaker, R.H. 1953. A consideration of climax theory: the climax as a population and a
621 pattern. Ecological Monographs 23:41-78.
622 Wieczorek, J., Guo, Q., and Hijmans, R.J. 2004. The point-radius method for georeferencing
623 locality descriptions and calculating associated uncertainty. International Journal for
624 Geographical Information Science 18:745-767.
625 Wisz, M.S., Hijmans, R.J., Peterson, A.T., Graham, C.H., Guisan, A., and NCEAS Predicting
626 Species Distributions Working Group. 2008. Effects of sample size on the performance of
627 species distribution models. Diversity and Distributions 14:763-773.

628 Young, B.E., Byers, E., Hammerson, G., Frances, A., Oliver, L., and Treher, A. 2016.
629 Guidelines for Using the NatureServe Climate Change Vulnerability Index, Version 3.02.
630 NatureServe, Arlington.
631 Zizka, A., ter Steege, H., Pessoa, M. de C.R, and Antonello, A. 2018. Finding needles in the
632 haystack: Where to look for rare species in the American tropics. Ecography 41:321-330.

633



634

635 **Figure 1**. The process of generating and analyzing a virtual species starting with generation of
636 the fundamental niche (step 1), projection to geographic space (2), and generation of omniscient
637 (3), precise (4), and imprecise (5) occurrences. Suitable climate area is analyzed within a
638 buffered region surrounding omniscient records (bottom row). The process was repeated 200
639 times for each combination of number of omniscient, precise, and imprecise records.

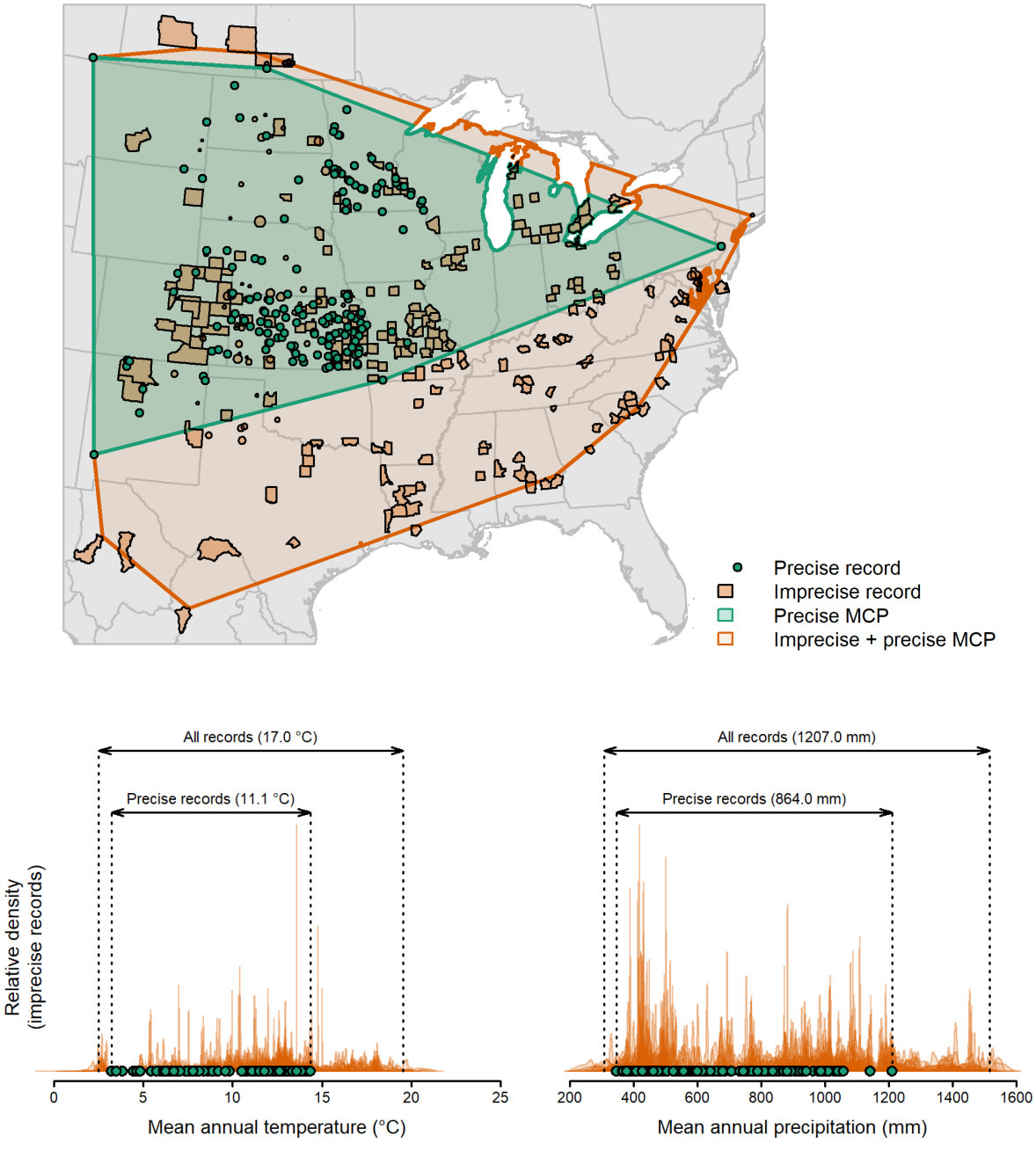


Figure 2. An example of including versus discarding imprecise records using *Asclepias viridiflora*. a) Extent of occurrence estimated using either just precise or precise plus imprecise records from the minimum convex polygon (MCP). Imprecise records are represented by the smallest geopolitical unit to which a record can be located, or by circles representing area of likely collection. Projection: Albers conic, equal-area. b) and c) Difference in climatic niche breadth. The values of temperature or precipitation at precise records are represented by points (green). The distributions of temperature or precipitation across all locations encompassed by imprecise records are represented by smoothed density kernels, one per record (orange). Niche breadths estimated using just precise versus precise plus imprecise are shown.

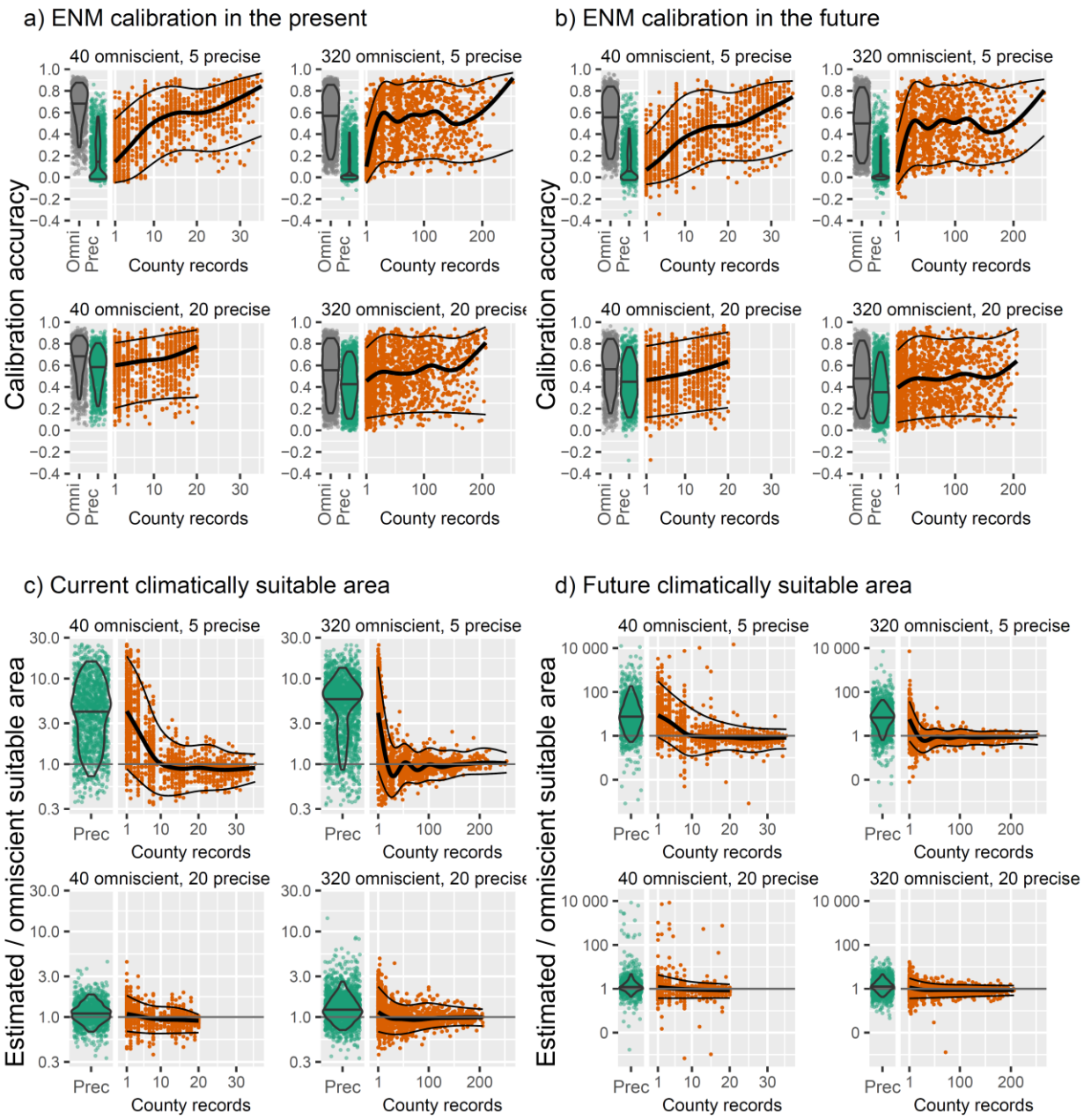
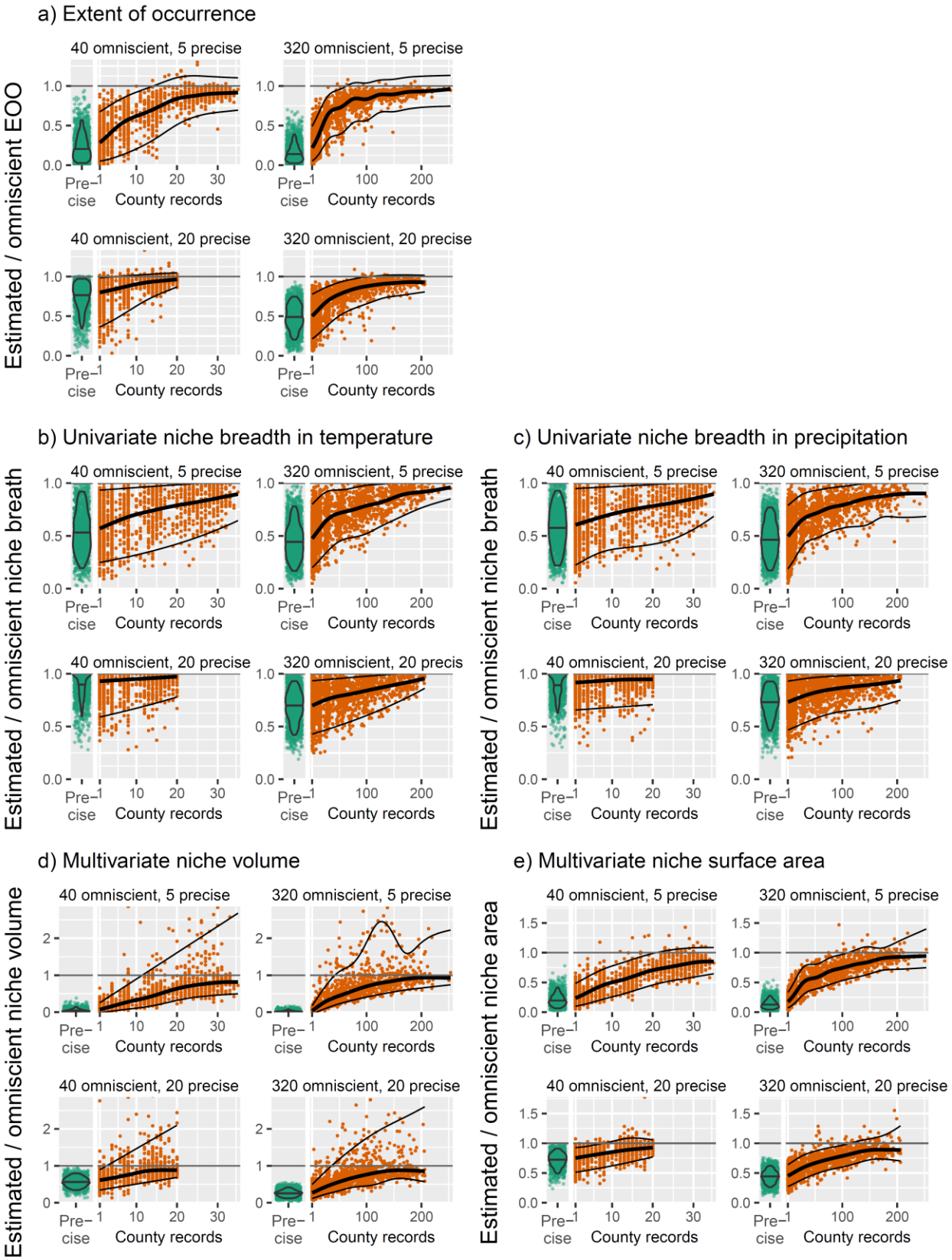


Figure 3. Adding imprecise records (here, reflected as the number of counties they occupy) improves niche model accuracy for the present (a) and future (b). Calibration accuracy is the correlation between the real probability of presence and model output. Current (c) and future (d) climatically suitable area are overestimated when only precise records are included. Values are the ratio between estimates using only precise or precise + imprecise records to estimates from models using all "omniscient" occurrences of a species. Note the log scales. Each point represents a species. Violins encompass the inner 90% quantile of values. Thick trendlines represent the median trend, and thin trendlines encompass the inner 90% quantile of values.



660

661 **Figure 4.** Adding imprecise records (here, reflected as the number of counties they occupy)
662 increases the accuracy of estimates of extent of occurrence (a), univariate niche breadth in mean
663 annual temperature (b) and total annual precipitation (c), and multivariate niche volume (d) and
664 surface area (e). In each case, values represent the ratio of estimates using only precise or precise
665 + imprecise records to estimates calculated using all "omniscient" occurrences of a species. Each
666 point represents a species. Violins encompass the inner 90% quantile of values. Thick trendlines
667 represent the median trend, and thin trendlines encompass the inner 90% quantile of values. To
668 aid visualization for univariate niche breadth and niche volume, the y-axis limits encompass only
669 the lower 99.9% of values.
670
671

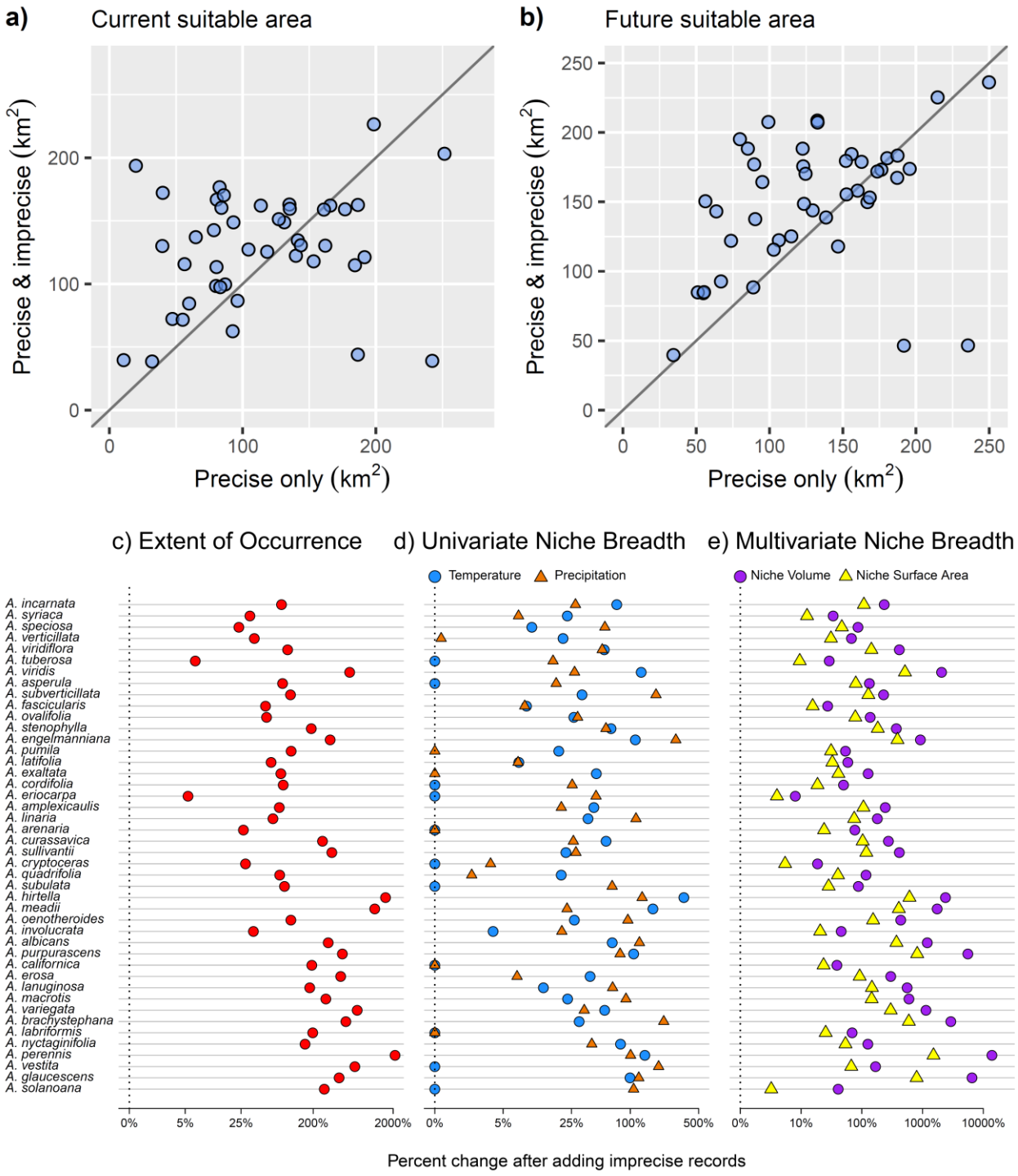


Figure 5. Effects of including imprecise records on niche breadth and climatically suitable area for *Asclepias*. Differences in climatically suitable area in the present (a) and future (b) under RCP8.5. (c) Change in extent of occurrence. (d) Change in univariate niche breadth in mean annual temperature and mean annual precipitation. (e) Change in multivariate niche volume and surface area of this volume. In panels c-e, species are sorted from top to bottom from most to least number of precise records. Also note the log scale along the x-axis.