Range models for all Land Plants

Cory Merow[1]

[1]Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06268, USA, cory.merow@gmail.com

February 28, 2024

Submission Type: Data Paper

Keywords:

Running Title:

Corresponding Author: Cory Merow

Abstract Word Count:

Text Word Count:

Reference Count:

Contents

Abstract

- The Botanical Information and Ecology Network (BIEN) is....
- We have developed a consistent, objective, and reproducible range modeling workflow ...species specific modeling decisions
- We have implemented these tools with the BIEN3 database (Botanical Information and Ecology Network) to model the ranges of ¿90,000 New World plants.
- Data preparation
- We describe all modeling decisions and emphasize where critical modeling decisions were made.
- We describe overall model performance to identify the strengths and weaknesses of the predicted ranges and baselines for future improvements.
- We conclude with an outlook on future plans for related range products at BIEN.

1 Introduction

why we need spatial biodiversity data

One of the fundamental units in ecology and biodiversity science is the geographic range of a species - the irregular region(s) on the globe where the species can be found. Despite the importance of knowing species ranges for basic ecology and for conservation, we have very limited knowledge of geographic ranges for the majority of organisms on the planet. Such ranges knowledge forms the baseline for understanding past, present and future processes influencing biological systems and may be critical for anticipating and managing species' responses to global change. The ranges of most of the 30,000 terrestrial vertebrates have been well studied and indeed are freely available (cite IUCN). These data have had enormous influence in basic research [3] and in conservation [5]. However, vertebrates represent less than 5% of the earth's species diversity, and less than 0.01 % of earth's biomass[cite Bar-on et. al. 2018]. For the majority of the other 95% of organisms (99.99% by biomass), no such comprehensive data are available. In plants, while there are isolated species range compilations - e.g, 600+ trees and shrubs in North America [6], 1000+ palms [7], 500 trees in Panama and Costa Rica [8] - ranges for the rest of the 380,000 (cite rarity paper) vascular plant species are largely lacking. The story is similar for many other groups, including organisms central to ecosystem functioning like insects, fungi, nematodes, and annelids, where essentially no systematic, standardized geographic range information is available. However, the absence of these ranges is not due to insufficient information in many cases. Indeed, modern analytic methods merging high-resolution environmental layers, existing biological inventory data, and statistical/machine learning techniques can often provide accurate range/species distribution models (SDMs) [15, 16]. To bridge the gap between biodiversity data and extensive knowledge of species' distributions, we developed a workflow that cleans and combines a variety of disparate data sources to provide comprehensive range predictions for New World land plants.

why building lots of ranges is hard

The major limitation in the production of species ranges for many organisms is an informatics challenge. Primary biodiversity data (PBD) are widely available in electronic format

(e.g. GBIF; `http://www.gbif.org/`) and many statistical/machine learning methods are widely available and often computationally efficient (e.g., [18, 19]). Instead, data limitations and computational challenges are apparent when scaling up to large ensembles of species. In short, it is infeasible for the modeller to manually assess all of the issues associated with the data utilized by SDM methods and check each step for plausibility in a workflow to generate 10s of thousands of SDM. SDM decision-making is more straightforward in the special case of one or two species [19-21], but difficult in the context of scaling up to 1000s or 100,000s of species when one must automate the process of making modeling decisions on a species-specific basis.

what we do here

Here we identify seven specific informatics scaling challenges and develop a novel informatics pipeline to utilize the increasing data deluge of biodiversity records to scale up range modeling to 10s of thousands of species with data preparation and modeling optimized on a species-by-species basis. In developing this workflow, we bridge fundamental data and computational barriers to predicting the geographic distribution of New World land plants which can be more generally applied to other taxa and regions. The culminating products are a set of present and future range predictions for 112,000 New World land plants, served openly through the BIEN database (`biendata.org`) for anyone to freely download and analyze, along with comprehensive metadata on the predictions and how they were obtained for each species.

mention future ranges and scenarios

2 Data Preparation

2.1 Primary Biodiversity Data (PBD)

Geographic ranges are computed from 'primary biodiversity data' [or PBD; 61]. PBD consist of taxonomic observations at a given time and location. PBD include collections in museums and herbaria, taxonomic monographs, ecological surveys, trait measurements,or descriptors of the community composition or environmental context of the sampling area

(e.g., survey or plot). PBD underlie the estimation of the species' ranges, the definition of vegetation types, and are the basis of our knowledge of phylogenetic diversity and distribution. Such data underlie any attempt to quantify patterns and determinants of diversity, and for predicting how ecological systems respond to global change [64].

Many types of PBD must be integrated to address the most pressing questions in ecology and global change biology. To access biological insights, it is necessary to integrate different sources of PBD across many institutions and researchers, as well as across diverse geographic, temporal and taxonomic scales. However, diversity data compiled across diverse sources and scales are fraught with multiple sources of error that can limit their utility if not properly addressed. As a result, efforts to understand the distribution of plant species and communities and to predict responses to global change necessitates a holistic approach to data cleaning, standardizing, and analysis. Such an approach must be conducted at a scale that is commensurate with the breadth of the questions being asked. Further, it requires identification, retrieval, and integration of diverse data from a global confederation of collaborating scientists across a broad range of disciplines to remove the barriers to developing geographic ranges for all species.

There are currently several efforts to compile different sources of PBD. For example, the Global Biodiversity Information Facility (GBIF), sequence data via Genbank, trait data via TRY, plot data via VEGBANK and SALVIAS, and citizen science obervations via iNaturalist and the USA NPN Nature's Notebook all offer enormous potential for assembling, aggregating and storing PBD. Nonetheless, integration of PBD either within or between these efforts will increasingly be limited because of errors in taxonomic names and geographic coordinates (Table 1). Further, datasets generated by individual researchers or teams are often geared toward specific questions. As a result integration of different data sources remains a challenge. As a result, data sharing and standardization remains difficult or currently impossible. Indeed, there remain significant challenges in mobilizing and integrating all PBD sources to either inform baseline estimates or derive information about biodiversity and ecosystem change [66].

2.2 The Botanical Information and Ecology Network (BIEN)

overview of BIEN

Usage of existing PBD presents many challenges when scaled up to many species. The wide availability of large amounts of PBD presents an opportunity to merge with species distribution models to move beyond ranges of a few percent of the earth's species to a majority of the earth's species. Building quality range models for 1,000s-100,000s of species is challenging because of errors in taxonomy, geography, and inclusion of nonnatural records. To develop range models for all New World land plants, we relied on data aggregated via the Botanical Information and Ecology Network (BIEN). The goal of BIEN is to assemble all the PBD (observation records from specimen collections, ecological plot surveys, and trait observations) for all New World land plants, from bryophytes to angiosperms. Version 3.0 of the PBD database (access via `BIEN3.org`), contains 200,000,000+ global observation records and over 380,000 (??) land plant species. [1] These 'scrubbed' data originally consisted of errors in ¿50% of records, including species names, geographic coordinates, and introduction of cultivars and exotics (Fig X). [2]

FiXme Note!

FiXme Note!

2.3 TNRS

FiXme Note!

[3] Occurrence records were cleaned to resolve taxonomic naming issues and remove records were latitude/longitude was not available or could not be verified using the Taxonomic Name Resolution Service (TNRS) (CITE). It functions as a webservice that compares a list of scientific names (Latin Binomials possibly followed by authorities) against a standardized list, checking for misspellings and synonyms. The code accessible via GitHub.

2.4 Geoscrubbing

FiXme Note!

[4] An increasing amount of PBD is being geocoded due to efforts by museums and an increasing interest in associating traits and gene sequences with specific locations. However, geocoding still has fairly high error rates. To identify incorrect (or potentially incorrect) coordinates, the BIEN workflow checks for (1) impossible coordinates (e.g. longitude greater than 180 in a WGS84 coordinate system); (2) suspicious coordinates (e.g. latitude or lon-

[1] FiXme Note: get numbers just for new world
[2] FiXme Note: add more stats from Brian M's ESA talk?
[3] FiXme Note: BRAD
[4] FiXme Note: Maitner

gitude exactly zero); (3) coordinates that fall in the ocean; (4) coordinates that are likely to reflect a political centroid; and (5) occurrences that fall outside of the lowest declared political division. Records that pass these checks are labelled as 'geovalid'. Political division boundaries were obtained from the Database of Global Administrative Areas (GADM; https://gadm.org), and included both current and historical boundaries of specified political divisions. Due to variation in political division names, both the declared political division names and the GADM political division names were corrected and standardized according to the GeoNames gazetteer (geonames.org) using the Geographic Name Resolution Service. Approximately 25 % of records, both in the New World and globally, were flagged as potentially erroneous in some way, while the remaining 75 % were labelled geovalid.

Native species resolver[5] Many observations are recorded of cultivated species. While this information can be useful (e.g., to study invasive species), it can also greatly distort range modelling when focus is on native distributions, as we consider here. The best way to filter cultivars remains an active research question. BIEN4.1 used searches for various key words in Darwin Core. Nonnative status was based on regional checklists ... BRAD .... Records that were cultivated or nonnative were removed, though native species lists were not available throughout the New World so this filtration was imperfect.

FiXme Note!

The BIEN Database

Brad:Any description of the specifics of BIEN 4.1 that you want to add

Once PBD are scrubbed and standardized, they are transferred into a common normalized database, BIENdb. BIENdb is unique among large biodiversity databases because it not only includes records of occurrence, but also of species traits and community/plot data (although these are each treated as occurrence data in the models described below). We have developed a schema for such a diverse set of PDB (see http://fs.vegpath.org/schemas/vegbien.ERD.pdf) These additional data types, in conjunction with phylogenetic relationships are a critical for making the leap from occurrence points to reliable range predictions to fill data gaps that otherwise might be apparent in occurrence data alone.

[5] FiXme Note: BRAD

3 Model Building

A challenge of scaling to 100s of thousands of species is in determining the appropriate modeling decisions for each species. Species vary enormously in the number of observations, the intensity with which they were sampled, their range size, the complexity of relationships with climate, etc. Obtaining reliable models is based on choosing appropriate settings, which rely on numerous decisions about data, ecology, and statistics that qualitatively and quantitatively affect all emergent predictions of ranges, biodiversity, or their past/future changes. Decisions on, for example, background selection, handling spatial autocorrelation, model complexity, and sample size are critical avoid to overfitting, underfitting, bias, and imprecision. The challenge remains to upscale best practices to many thousands of species and automate the decision-making process to produce robust models. Automated decision making in our workflow is a fundamentally new approach to studying geographic patterns in biodiversity, detailed in Fig. 2.

Model algorithms. Three different range estimation methods were used depending upon the sample size of presence locations. For species with 10 or more records, inhomogeneous Poisson point process (PPM) models were used. (Warton and Shepherd, 2010; Renner et al., 2015). PPMs are a generalization of the commonly used MaxEnt algorithm (Phillips et al., 2006) that allowed us the flexibility to rely on established modelling tools while optimizing them with customized new approaches described below (Elith et al., 2011; Merow et al., 2013). Model settings were chosen to balance overfitting (under estimating range sizes) with underfitting (excessively smooth models that over predict range size). Ranges for species with 3-9 records were built using a range bagging algorithm (Drake, 2015; Drake and Richards, 2017). A species with a 1-3 records were assigned a range that included only the $100km^2$ cell(s) where it was found.

3.1 Poisson Point Process Models

3.1.1 Processing Presence Data

A variety of steps are taken to prepare occurrence data after the cleaning steps but before modeling. If multiple records were found in a 10km grid cell, only one was retained. To reduce

spatial autocorrelation, we thinned occurrences to ensure at all retained records were at least 20km from one another using a single random sample of from the 'spthin' (Aiello-Lammens et al., 2015) default thinning algorithm. If fewer than 20 records were available, we did not apply spatial thinning in order to avoid reducing sample size. A maximum of 20,000 records were retained ((randomly) per species to avoid excessive run times under extensive sampling.

Outliers in geographic and environmental space were determined based on a Grubb's outlier test with p=1e-5 (**?**) implemented with the R package 'outliers' (**?**). For each test, we calculated the centroid of all records in geographic or environmental space (respectively) and then the distance from each point to the centroid. The one-sided Grubb's test then determines whether the single largest distance is an outlier relative to all other points. If it was determined an outlier, the point was discarded, and the test repeated on the remaining points until no points were determined outliers. The small p-value was chosen heuristically to ensure that only fairly blatantly outlying points were removed. Supplementary Figure XX shows some representative graphical examples of outliers.

Finally, presences were clustered into five folds for later cross-validation using an algorithm newly developed for this workflow. Folds were spatially stratified to minimize extrapolation during modeling evaluation. Folds were generated by first computing a k-means cluster on the coordinates of records, seeking 25 clusters. These 25 spatial clusters were then randomly assigned to five folds. In contrast to directly computing five clusters, this reduces the chances that a significant portion of environmental space was withheld from model training and hence reducing artefacts from extrapolation (Phillips, 2008). These folds may have unequal sample sizes depending on the spatial clustering of records, hence this clustering reduces the influence of spatial autocorrelation on evaluation statistics. Since this fold assignment was the only stochastic component in the modeling workflow we set a random seed based on converting each species' name to an integer to ensure that all results are exactly reproducible.

that i remember - Xiao?

3.2 Processing Environmental Data

Environmental covariates included both climate and soil layers. Climate covariates were obtained from WorldClim 2.0 at 10 km resolution (Hijmans et al., 2005). Predictors included four bioclim variables - mean annual temperature, mean diurnal temperature range, annual

precipitation, precipitation seasonality - as well as precipitation in warmest quarter/ (precipitation in warmest quarter + precipitation in coldest quarter) , and aridity . The bioclim variable were chosen based on correlations across the New World of r ¡0.7. The additional predictors were added based on expert recommendations to capture seasonality in tropical climates and [PATRICK] and because they also had r¡0.7 with the bioclim predictors. We also included soil layers from `soilgrids.org` - depth to bedrock, proportion clay and proportion silt in the first four soil horizons, and mean bulk density and mean pH in the first four soil horizons. These were chosen as the largest subset of soil layers hypothesized to be relevant for large scale biogeography patterns which also preserved r¡0.7 with the climate layers. To generate these layers, we aggregated the 250m resolution layers available on `soilgrids.org` to the 10km grid defining the climate layers. This set of 11 covariates was the starting point for each species; within each species-specific modeling domain we further subsetted predictors to ensure they locally had r¡0.7 by retaining the largest subset of predictors below this correlation. Our emphasis on removing correlated predictors is based on reducing the amount of correlation on which fitted models may depend to reduce the influence of changing correlations among covariate in the future forecasted ranges.

McGill - please provide a ref

Patrick how was this calculated, and please add ref (accumulated aridity based on the max accumulated water deficit)

The main variable that is known biologically important but not captured in BIOCLIM - indeed I know somebody who did a formal analysis and it was the first PCA of monthly data that didn't get picked up by BIOCLIM is whether the precipitation peak is in the warm or cold period. This is for example the key driver of praries vs Mediterranean climate, Northwest Evergreen vs NE temperate, Chihuahuan vs Sonoran vs Mojave deserts etc.

To account for the effect of dry season length and the water deficit experienced by vegetation during dry periods we derived and accumulated aridity index that identifies the maximum duration and accumulated water deficit for consecutive months where potential evapotranspiration exceeds mean monthly precipitation. The accumulated aridity index was created through the following steps:

Mean monthly potential evapotranspiration (PET) was calculated for each month in both baseline climate and in projected future climates. Calculation followed the methods of the CGIAR-CSI Global-Aridity and Global-PET Database (http://www.cgiar-csi.org) PET = 0.0023 * RA * (Tmean+17.8)* ($TD^0.5$)$RA = monthlytotalextraterrestrialradiation(datafromCGIAR)Tmea$

$mean monthly temperature TD = mean monthly diurnal temperature range Runs of consecutive months where P$
$monthly precipitation were identified for each pixel The accumulated aridity index is the sum of the month aridity (p$
$precipitation$)

Figure 17. Accumulated aridity index for baseline climate. Color ramp shows Maximum accumulated aridity for consecutive months of PET ¿ mean monthly precipitation. Red = highly arid; blue = no water limited months.

Formula for monthly PET is here:

http://www.cgiar-csi.org/wp-content/uploads/2012/11/Global-Aridity-and-Global-PET-Methodology.pdf

I used worldclim monthly data to produce the projected future aridity stuff. The accumulated aridity over a run of consecutive water limited months that we used for models was my idea (I think) – to address concerns of our Asia collaborators that dry season length wasn't being captured in the standard bioclimatic variables.

Looks like there's an updated baseline product here (Robert Zomer is a collaborator in Asia):

https://cgiarcsi.community/2019/01/24/global-aridity-index-and-potential-evapotranspiration-climate-database-v2/

Domain selection.

- fit and projected based on occupied ecoregions and their neighbors

2e4 background

Algorithm settings

PPM / Maxent models were fit using regularized down-weighted Poisson regression (Renner et al., 2015) fit with the R package glmnet (**?**). Different feature classes were added depending on sample size: linear and quadratic (all species), product (species with ¿100 records), and hinge (species with ¿200 records). On occasion, when models failed to converge with these feature class settings, the next simpler feature set in this hierarchy was attempted. In cases where linear and quadratic features failed, we used only a subset of the candidate predictors to generate linear and quadratic features, selecting the five (or three, if five fail) predictors with the highest univariate correlations with presence/background data.

This sequential approach of applying simpler features to failed models made our workflow very robust to idiosyncrasies in each species' data set, ensuring that models could be fit in all cases. Notably, we chose only to consider linear and quadratic features paired together in order to reflect the assumption that species' responses to environment should often be modal; while we did not force modal responses, we always included them as an option during feature selection.

The regularization parameter was determined based on 5-fold cross-validation with each fold, choosing a value one standard deviation below the minimum deviance (Hastie et al., 2009) (also a standard choice built into the cv.glmnet function). This approach allowed us to find an 'optimal' (in the sense of balancing overfitting with underfitting) regularization parameter based on efficient computation of the entire regularization path (**?**). This resulted in five models per species which were then combined in an unweighted ensemble. This ensemble prediction can be interpreted as a relative occurrence rate (sums to 1 over the modeling domain), such that it predicts the relative probability that an observed presence came from each cell in the domain (Fithian and Hastie, 2013; Merow et al., 2013).

Projection

- only project to fitting domain
- project into future in 2050, 2070 for rcp 2.6 and rcp8.5
- When models were projected into the future, we limited extrapolation to 1 standard deviation beyond the data range of presence locations for each predictor. This decision balances a small amount of extrapolation based on patterns in a species niche with limiting the influence of monotonically increasing marginal responses, which can lead to statistically unsupported (and likely biologically unrealistic) responses to climate.
- The five models from each fold were then combined in a weighted ensemble, with weights determined by the partial AUC calculated between sensitivity between 0.8 and 0.95. We chose partial AUC because we are uninterested in thresholds that result in lower than 80% sensitivity for SDM applications. The upper limit of 0.95 was chosen because 5% omission rate was used for creating binary maps (to stack and estiamte species

richness), and hence higher sensitivity values were not relevant for our application.

Model evaluation.

- tons of options
- test AUC now
- other metrics are highly correlated with this, but probably better representation of performance
- like to move toward $pAUC_{.8-95}$

Modeled Products

1. continuous ROR rasters
2. ROR uncertainty
3. binary shapefiles. 5% training pres. Continuous predictions of relative occurrence rate were converted to binary presence/absence predictions by choosing a threshold based on the 5^{th} percentile of training presence locations.
4. binary rasters
5. trinary maps. pAUC. upper and lower bound on range BASED ONLY ON THRESHOLD
6. model performance statistics
7. model metadata
    - standardized
    - currently a short list of input and output stats

3.2.1 Range bagging

Domain selection was based on the same rules as for PPMs; occupied ecoregions and their neighbors. **Algorithm settings**

- Only linear, quadratic, and product features optimal regularization determined on a species-by species basis using spatially stratified cross validation. find optimal regularization for each fold

- formulas: linear, quadratic, product, depending on sample size

- some formulas break, default to next simplest formula

Projection

Modeled Products

3.2.2 Points

Modeled Products

3.3 High performance computing

A HPC workflow allows for running 110k+ species range models in an HPC environment with necessary job management and error recovery. Scaling up to produce high quality range models for 100,000+ species poses two major types of challenges. First, running a single SDM for one species is moderately computational intense (e.g. runtime of one minute). But to scale this to 100,000+ species requires moving to an HPC environment, which is prohibitively complex for most biologists. We implement a pipeline that uses PBD from the database described above (i), iterates over species, runs appropriate SDMs for each species, and then captures output as species ranges to be transferred back to the database schema (using spatial extensions to RDBMS such as PostGIS).

4 Modeled Products

- continuous ROR rasters

- ROR uncertainty

- binary shapefiles

- binary rasters

- trinary maps
- model performance statistics
- model metadata

5 Summary statistics

1. any informative way to break down how species are split along the workflow? or too trivial
2. distribution of species over decision tree bins
3. variation in our performance metrics - how many species are we doing ok with?
4.

6 Discussion

overview Biological inventory data are large in scope, growing, and increasingly organized in electronic databases. While the datasets needed to generate geographic range information will continue to grow over time, we believe the scientific community is in a good position to now begin addressing these questions about geographic ranges. As already noted, SDM is a widely accepted and used tool but with the complexities of assembling climate layers, applying appropriate expertise in modeling techniques, and computational costs of running SDM, most studies address a handful or a few dozen species at a time. The key challenge to producing such geographic ranges for many taxa over many time periods is not driven by limitations in availability of biological inventories. Instead, the key limiting factor is the lack of an easy to use computational pipeline for the scrubbing and standardization of data followed by the mass production of robust geographic ranges. The products of our workflow provide an enormous, standardized resource for the botanical and biodiversity research communities. By virtue of producing the first robust models for most land plants, these ranges can be treated as data for subsequent studies in global change, botany, evolution, functional ecology, biogeography, and macroecology. Producing geographic ranges for

many groups of organisms enable addressing a wide variety of questions (SUCH AS?).

no models failed using these settings.

interpreting maps

potential applications

caveats Automating model building for 110kspecies is not without flaws and some caveats should be recognized. Notably, sample size remains small for the vast majority of species,(52k species with ¡ 9 unique presences, 30k with 3-9 presences and 30k with 1-2 presences) hence many ranges are surely estimated incompletely. It is impossible to automatically detect all problematic, outlying, or nonnatural occurrence records and those that remain may influence range predictions. Given our attempts to avoid overfitting, the species distribution models are more likely to underfit spatial distribution patterns and consequently may predict ranges larger that those realized for some species. That is, the models may predict suitable habitat in locations that are inaccessible to the species (but in in similar environmental conditions to where they occur) or predict suitable habitat slightly beyond realized range edges due to fitting relatively smoothed response curves. To offset this, cells where presence was predicted by Maxent further than 1000km from any presence record were removed from the range. Correction has not been made to account for variation in sampling effort or detection probability. Like any range map, our predictions represent hypotheses about spatial occurrence patterns. In spite of these caveats, predictions for the vast majority of species are reliable and are well-suited for macroecological analyses.

Our range modeling efforts are a dynamic enterprise and we are constantly exploring ways to improve predictions, leading to periodic updates in our database. Planned updates include choosing optimal models settings tuned specifically for each species, accounting for sampling variation, and improving occurrence data cleaning methods. We will employ version control to maintain accessibility of all past versions as updates are released.

1. These are (or will be) spatial models, not niches

2. suitable for macroecology

3. large percentage pretty darn good
4. total information gain from inter/extrapolation. what regions do we learn the most from
5. where is our performance best?
6. touch on where we stand to learn the most, but don't scoop the sampling paper
7. code available on github - maybe nathan can simplify to a more portable version
8. how this works with HPC
9. compare to other mapping initiatives? california? loks for specific taxonomic groups and see how we compare. we just need to do a similar job to people who've focused in detail on a specific group
10. how does this compare to jsut the point data?
11.

6.1 Serving models

1. RBIEN
2. Output types
3. shrink files

1 paragraph overview of RBIEN The BIEN package for R [cite R] allows users to maitner query the multiple data types within the BIEN database. This package converts user input into PostgreSQL queries that are submitted to the BIEN database, and then returns the resulting output. Function names follow a three part naming structure consisting of : (1) the name of the package ("BIEN_..."), (2) the type of data accessed by the function, (e.g. "...ranges_..." accesses range maps), and (3) how the data are being queried (e.g. "...species" functions return data for a given species). Thus, the function BIEN_ranges_species() returns range maps for a given species or set of species. Currently, the BIEN package can retrieve the

follow data types: occurrence records, phylogenies, plot data, range maps, species lists, stem data, taxonomic information, trait data, and metadata on the BIEN database itself. Species range maps can be queried both taxonomically (e.g. for a given species or genus) and spatially (e.g. range maps that intersect a polygon, bounding box, or focal species' range). Once a desired set of data are downloaded, attribution information for use in resulting publications can be generated using the function BIEN_metadata_citation().

7 Future Directions

Our range modeling efforts are a dynamic enterprise and we are constantly exploring ways to improve predictions, leading to periodic updates in our database. We will employ version control to maintain accessibility of all past versions as updates are released.

Planned methodological updates include continually improving occurrence data cleaning methods, examining a variety of methods to account for sampling variation, comparing predictions based on different environmental data sources, and comparing predictions from different algorithms tuned for poorly sampled species.

extending predictions globally

Automating model building for 110k
baseline for distributions
ongoing advances and new syntheses as they become available

8 Acknowledgements

C.M. acknowledges funding from NSF Grant DBI-1913673 and NSF grant DBI-1661510.

References

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. 38(5):541–545.

Drake, J. and Richards, R. (2017). Estimating environmental suitability. pages 1–29.

Drake, J. M. (2015). Range bagging: a new method for ecological niche modelling from presence-only data. *Journal of The Royal Society Interface*, 12(107):20150086–9.

Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., and Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17:43–57.

Fithian, W. and Hastie, T. (2013). Finite-Sample Equivalence in Statistical Models for Presence-Only Data. *The Annals of Applied Statistics*, 7(4):1917–1939.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, New York, 2nd edition.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978.

Merow, C., Smith, M. J., and Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069.

Phillips, S. J. (2008). Transferability, sample selection bias and background data in presence only modelling: a response to Peterson et al.(2007). 31(2):272–278.

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

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

496 Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-
497 absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*,
498 4(3):1383–1402.

499 9 Figures

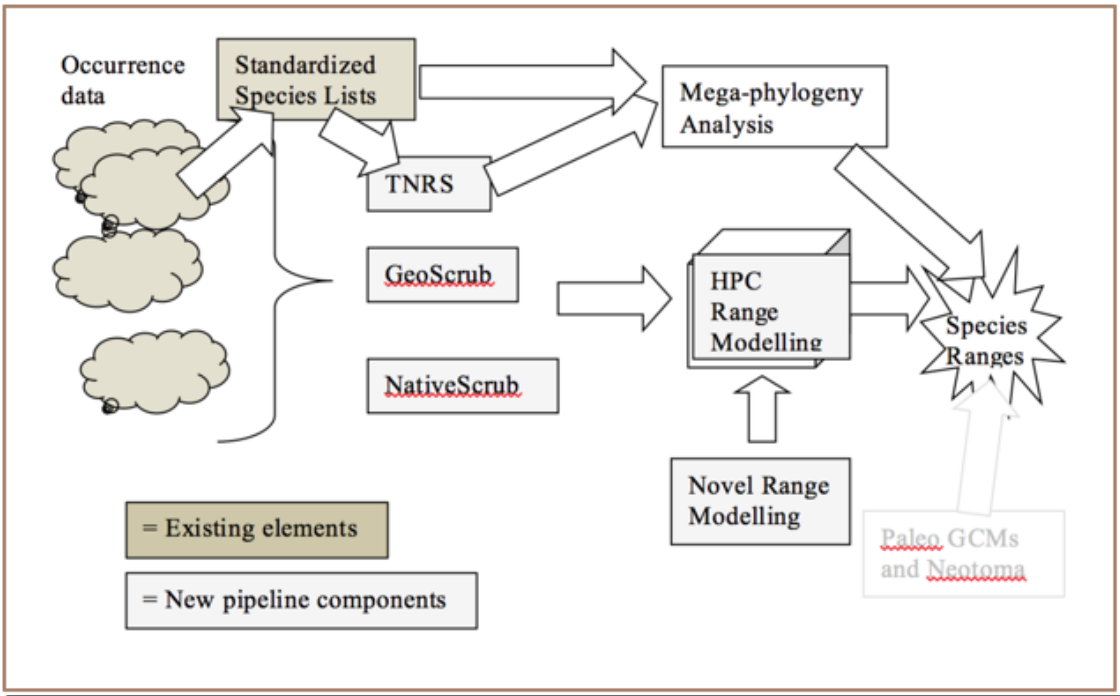

Fig. 1. Generic computational pipeline to clean biodiversity data and standardize implementation of analyses fundamental to biodiversity research.

Figure 1: The BIEN range modeling workflow. To be updated. please send the .ppt.

10 Online Appendix