Methods - sdms for restoration

2 MATERIALS AND METHODS

2.1 Study system

The 355 modelled species were selected by land managers in arid & semi-arid areas Western North America and are species they already or immediately intend to use in restoration. The spatial domain of modelling broadly encompasses the Western United States being bounded by the Pacific Ocean, the 50th parallel at north, -100 degrees at East and Mexico to the south. This domain has tremendous variation in amounts of elevation, temperature, and precipitation.

Species Occurrence Records

Species presence records were collected from Natural History Museums, citizen science initiatives, and standardized ecological monitoring programmes using R (@chamberlain2024rgbif, @maitner2023bien, @michonneau2023idigbio). These records were filtered to only those collected after 1950, with coordinate uncertainty less than 250 meters, and only the most recent record per 90m cell was retained. All of these records were then manually reviewed by species, where records with more 2 or more variables in the 2.5% quantiles of several environmental variables (bio1, 4, 10, 12, & 19; TABLE X) and distance were flagged. During subjective manual review all digitized herbarium specimens which were suspect were reviewed, while other records were dropped based on the analysts review of other occurrence records.

Species absences were generated using three processes, and we sought to have a 1:1 ratio of presences to absences. True absences were acquired from a massive ecological monitoring program Assess, Inventory, and Monitor (AIM), these absences accounted for a percent equivalent to land managed by BLM within the species range (e.g. if 20% of land ownership across a species range was BLM, than the number of presences * 0.2 defined the number of these absences). Likely absences, representing 15% of records, were generated outside the known range of the species, but bounded to be within 50 miles of the extent extent of occurrences. Pseudo-absences (PA) were randomly selected from areas in, or within XX km, of the species range but greater than 10km from an occurrence, these records accounted for ((1 - (% BLM Land + 15 % LA records)) * 1.25). Because most of these species are highly abundant in order to reduce the probability that a PA was drawn within a species the environmental niche linear discriminant analysis (LDA) was used. The LDA utilized the presences, true absences and likely absences as the classes and several environmental variables (@venables2002mass), identified by vifstep with theta = 10 as displaying at most moderate collinearity, as independent variables (@naimi2014usdm). As many records, classified by the LDA as originating from the Presence data set, from the pseudo-absences could be removed as to achieve a class balance between presences and absences.

Environmental Variables

XX variables were used in generating the species distribution models (table X, @karger2017climatologies, @hengl2017soilgrids250m, @ivushkin2019global, @tuanmu2014global, @yamazaki2017merit, @amatulli2020geomorpho90m, @sanderson2002human). These variables were selected based on authors previous work, and represent variables commonly associated with the empirical distributions of taxa in arid and semi-arid regions. All variables were downloaded from source and re-sampled to 90m resolution using terra (@hijman2023terra).

Species Distribution Modelling

Random Forests models were generated for each species following several steps to reduce the number of features. While superfluous features generally do not notably decrease the performance of random forests, they increase the amount of time required to predict models onto gridded surfaces. The Boruta algorithm was used on the full stack of 44?? independent variables to remove un-informative variables @kursa2010boruta), and the variable importance factor (VIF) scores were then used to subset the most informative features in several auto-correlated pairs, reducing the independent variables to at most 33.
Recursive Feature Elimination (rfe) was then used to identify the fewest number of independent variables which could either increase model performance (measured using accuracy), or only decrease it by 1.5% relative to the full set of variables, these variables were subset and used as features for random forest modelling with all default values except for optimized mtrys (@kuhn2008caret, @liaw2002randomForest). Models were then predicted onto gridded surfaces which exceeded the species range by 50 miles using terra's predict.

Patch identification

In order to identify patches which may support populations raster values less than 0.6 were masked as NA's and a rooks case search was conducted (@hijman2023terra).

To determine whether habitat suitability could predict species abundance mean weighted patch suitability $PS = \sum_{i=1}^{n} x_i$ was regressed against abundance using GLM... All AIM plots were queried for all occasions which a target species was observed via either the Species Richness, or Line-Point Intercept method (150 points along three equiangular transects over a 25m radius outside a 5m radius buffer zone). Occurrence of a species only along SR was considered 1% cover, while the canopy cover from LPI was used.

To determine whether connectivity could predict the presence of a population in a patch, connectivity was calculated between every patch, and jacknife-resampled for logistic regression where connectivity scaled from 0 to 100 served as a the independent variable.