Seeds of Change application
TECHNICAL DOCUMENTATION
version 0.2.0, April 16, 2024

Matthew M. Kling[1,2,†] and Patrick Gonzalez[3,4]

[1]Department of Biology, University of Vermont
[2]Gund Institute for Environment, University of Vermont
[3]Department of Environmental Science, Policy, and Management, University of California, Berkeley
[4]Institute for Parks, People, and Biodiversity, University of California, Berkeley
[†]Contact email: mattkling@berkeley.edu

ABSTRACT

Seeds of Change is a web-based application designed to help inform decisions about seed collection and planting locations for ecological restoration projects, by analyzing and visualizing spatial patterns of five climate and 16 soil variables across the geographic ranges of 5211 plant species in California. The tool incorporates data on current plant occurrences and climate and projections of potential future climate change. While other seed provenancing decision support tools have been developed, *Seeds of Change* includes a set of features not previously offered, including data on soils, data on climate variables ecologically important to California plant species, models of plant species distributions, and methods for exploring alternative assumptions about local genetic adaptation patterns. This document provides background, usage instructions, and technical details.

ACKNOWLEDGMENTS

Contents

1 INTRODUCTION

1.1 Background

The *Seeds of Change* tool is designed to help inform ecological management activities including matching the environments of seed provenancing and planting sites under projected climate change, collecting material from diverse environments for seed banking or restoration efforts, and assisting gene flow in the face of climate change. The tool integrates data on current plant species distributions, past and projected future climate, and soil characteristics.

The core function of *Seeds of Change* is to measure the environmental similarity between a focal site (either a location where plant material was collected, or where it will be planted) and other sites across the range of a given plant species, typically spanning two points in time. It also performs environment-based cluster analyses to estimate seed zones within a species' range. Users can explore how these results change under different model settings, such as alternative climate change scenarios, the relative importance of soil versus climate to a species, and the spatial scale of population genetic adaptation to the local environment.

The tool uses data on five ecologically important climate variables, including total annual precipitation, summer maximum temperature, winter minimum temperature, total annual climatic water deficit, and total annual actual evapotranspiration. It also uses five summary soil variables derived from 16 soil variables of chemical (e.g. pH, cation exchange capacity, and carbon content) and physical (e.g. sand, silt, clay, and coarse fragment content) properties at various soil depths.

The basic rationale for localized environmental information being relevant to management questions like these is the assumption that plant populations are genetically adapted to the historic environment of the locations where they live. Research suggests that populations often, but by no means always, exhibit true "local adaptation," meaning they exhibit better performance in their local environments than in other environments across the species range. Even in the absence of proper local adaptation, plants of course have demonstrated an ability to survive, on timescales that vary, in the locations where they are found. This implies that understanding patterns of climate departure from these historic conditions can help to inform climate change vulnerability assessments.

Estimating the environment that a plant population inhabits and to which it may be adapted requires a definition of "population." Biologically, the spatial scale of populations will depend on the rates of seed and pollen movement among nearby sites, relative to the rate of natural selection within a site: populations with strong selection and limited movement may be highly adapted to their local microenvironment, whereas populations with substantial seed and pollen movement relative to local selection may instead be adapted to the average environment across a larger swath of the species range. This tool uses spatial "smoothing" filters to approximate different scales of local genetic adaptation.

1.2 Use cases

This tool can be used to inform a variety of interrelated management activities, including:

1.2.1   Predictive provenancing The goal of predictive provenancing is to choose locations, across a species range, to collect seeds to plant in a given restoration site, often by analyzing spatial environmental patterns to estimate which populations may be relatively more adapted to the environment of the planting site. In the *Seeds of Change* tool, setting the focal location to the restoration site, the focal activity to "planting," and the display variable to "environmental difference" will visualize how similar the historic environments of sites across a species range are to the future environment of the restoration site. Sites with more similar environments are hypothesized to be more suitable as seed collection locations.

1.2.2   Predictive planting The inverse of predictive provenancing, the goal of predictive planting is to choose locations to plant seeds that have been collected from a particular site. Setting the focal location to the known collection site, the focal activity to "seed collection," and the display variable to "environmental difference" will visualize how similar the future environments of sites across a species range are to the historic environment of the collection site. Sites with more similar environments are hypothesized to be more suitable as planting locations.

1.2.3   Assisted gene flow and assisted migration These are two practices that require careful consideration of environmental, ethical, and legal issues before implementation (*1*) (*2*). Managed relocation practices involve helping species overcome dispersal limitation in the face of climate change-driven fitness declines, by moving plant material to more suitable sites within or beyond their current ranges, respectively. They involve similar data considerations to the above two practices, but with a broader species-level conservation focus. The *Seeds of Change* tool can help inform these activities by matching seed collection and planting sites within a species range, and by using the scatter plot to understand predicted climate change at a given focal site in the context of climate variation across the species entire California range, which could be an indication of the species-level climatic niche. Assisting gene flow involves managing both inbound dispersal of beneficial alleles into a given population that needs rescuing, and outbound dispersal of unique alleles that need rescuing; these cases are addressed by the "seed collection" and "planting" modes in the tool, respectively. Assisting migration involves outbound relocation of a species to sites outside its current range; while controversial due to the potential ecological risks it carries, it can be explored by deselecting the "limit results to species range" option when the activity is set to "seed collection."

1.2.4   Climate vulnerability analysis The app can be used in multiple ways go get insight into which parts of a species range may be more vulnerable to projected future climate change. It can display maps of climate change exposure, i.e. the magnitutde of climate change at each site across the region, for individual climate variables. It also includes a species-specific "multivariate climate change" metric that combines the magnitude of projected change in all five climate variables in each grid cell, normalized by the realized niche breadth of a species in California. For individual focal sites, the default "environmental difference" output variable provides insight into vulnerability when activity is set to "planting"; sites with higher vulnerability will have fewer and/or more distant areas with low environmental difference values. Finally, the scatter plot also illustrates climate change at the focal site (arrows) relative to climate variation across the species range (points), which can be used to visualize whether a given site is projected to remain within the realized climatic niche of a species as climate changes or is projected to become unlike any conditions the species currerntly inhabits in California.

1.2.5   Admixture provenancing and seed banking   A separate category of seed collection strategies aims to source seeds from populations across the species range that encompass a wide diversity of different environments, with the goal of maximizing the diversity of adaptive genetic variation that is sampled. The predictive "seed zones" output is designed to support these efforts, by running an environmental cluster analysis on sites across the species range to identify zones that are as internally homogeneous, and as distinct from other zones, as possible. Clustering can be done using past or future climate and/or soil data, and with varying scales of spatial smoothing applied.

1.2.6   Stratified sampling   Beyond informing seed collection efforts, the tool can also be used to inform other activities that require stratifying by or controlling for environmental variation across a species range, such as establishing sites for field surveys or genetic sampling.

1.3   Limitations

All outputs are model estimates with inherent uncertainty, and should be employed at the user's own risk, with the knowledge that they are hypotheses that may not reflect true plant performance patterns. Limitations include:

1.3.1   Environmental data are uncertain   Climate and soil data for the historic and future time periods will not perfectly reflect conditions on the ground, due to a combination of factors including uncertainty in the data interpolation process, scale differences between the 1 km resolution data and the microenvironment experienced by a plant, and uncertainties about future climate change. In addition, while the particular variables used here were selected for their general relevance to California plant ecology, they may not be the maximally relevant variable set for any individual species.

1.3.2   Species ranges are only estimates   While the species range maps reflect best efforts to model species distributions by associating known species occurrence localities with factors like climate and landscape intactness, true species distributions are almost never known. Furthermore, only the California portion of species ranges are included in the application, which, for non-endemic species, artificially limits the set of potentially relevant provenancing or planting sites and underestimates realized niche breadths.

1.3.3   Adaptive patterns are unknown   While the 'smoothing' setting in this tool is designed to let users explore potential patterns of local adaptation based on different assumed spatial scales of homogenizing gene flow, true adaptive genetic patterns are far more complex and are unknown for most species.

2 USAGE INSTRUCTIONS

2.1 Input settings

The model settings are grouped into a "species" tab that includes parameters related to the species biology and a "context" tab that includes settings related to the broader analysis. These options can be set based on the particulars of your species and use case or they can be changed iteratively to explore the sensitivity of outputs to these choices. Each time a setting is changed, results are recomputed and displayed on the map and scatter plot in the results panel.

2.1.1   Species To choose a species, type its scientific name in the species field, and select from the dropdown menu listing taxa with names similar to the text you've entered. This will load a map of the species' estimated geographic range within California, modeled at 1 km spatial resolution. The tool compares the environments of every grid cell across this range.

Alternatively, you can enter "NONE" in the species box. This will yield a species-agnostic analysis covering all of California, which could be useful in cases when the focal species is not known or its distribution is not well reflected by the model.

2.1.2   Site This tool compares the environment of each grid cell across the species range to the environment of a selected "focal site." To select a focal site, click a location on the map, or enter a latitude and longitude in the input box. A red marker will appear on the map, indicating the location of the selected site. The environment at this location is extracted from gridded climate and soil datasets with the same 1 km spatial resolution as the species range map.

2.1.3   Activity Next, choose whether the focal site is a "seed planting" or "seed collection" site. If "planting" is selected, the predicted future environment at the focal site will be compared to the smoothed historic environments of sites across the species range. If "collection" is selected, the smoothed historic environment of the focal site will be compared to the predicted future environments of sites across the species range.

The tool assumes that seeds can only be collected within a species' range, but does not restrict planting sites to the range. If the activity is set to "planting," a focal site anywhere in the state can be selected. If the focal site activity is set to "collection," the site must be within the species modeled range; by default, results are shown only for prospective planting sites within the range, but setting "limit results to species range" to "FALSE" will show results for all of California.

2.1.4   Timeframe & emissions scenario Use the drop-down menus to select the time period and greenhouse gas emission scenario, which will load future climate change projections based on an ensemble of general circulation models (GCMs). The tool includes data for three future time periods, for each of three different greenhouse gas emission scenarios ("Shared Socio-economic Pathways", or SSPs, defined by the Intergovernmental Panel on Climate Change). SSP5-8.5 is the highest-emissions scenario, while SSP1-2.6 assumes aggressive reductions in global emissions. For a given SSP, results are based on the 30-year

mean climate, averaging across five different GCMs (the individual models are depicted with arrows in the scatterplot as discussed below but the ensemble mean is used in calculations).

These future projections are compared to historic climate from the period 1981–2010, with one time period used for the focal site and the other for sites across the species range, as discussed above. It is also possible to select the historic time period as the focal period, in which case, environmental comparisons are made across the species range using only baseline data without considering projected future climate change.

Note that soil data do not change by time period or emissions scenario.

2.1.5 *Species distribution model (SDM) threshold* The species range maps in this tool are based on models that estimate continuous gradients of occurrence likelihood, in which every species has suitability values ranging from 0 to 100 across the state. These are not probabilities; they represent occurrence scores as a percentage of the highest occurrence score for the species anywhere in the state. This tool "thresholds" these suitability maps, displaying only those areas with a modeled occurrence score higher than the value set by the "SDM threshold" slider. Choose a lower value to select a broader range (which will likely include more actual populations, but probably also more places the species doesn't actually exist). Choose a higher value to select a narrower range (which will reduce the inclusion of places the species doesn't actually exist, but may also omit more true populations).

2.1.6 *Smoothing radius* The "smoothing radius" setting allows users to explore different assumptions about the spatial scale of local genetic adaptation across populations. This tool uses climate and soil data with a 1 km spatial resolution. To represent a given scale of adaptation, each historic environmental layer is "smoothed", replacing the value of every grid cell with the weighted average value of all the grid cells within a given distance defined by the "smoothing radius" slider (the species suitability scores described above are used as weights during averaging). The smoothed layer serves as an estimate of the environment in which the population in a given site evolved. (Future environmental layers are not smoothed, because they represent the climate that the planted individuals will experience, which is not subject to evolution.)

Patterns of genetic variation emerge from a complex balance between natural selection (adaptation), dispersal (gene flow driven by movement of seeds and pollen), genetic drift (random loss of genetic variation, primarily due to low population sizes), and mutation (which adds genetic variation but is relatively rare). In species with high rates of natural selection relative to the other processes, populations are more likely to be adapted to the highly localized environment where they occur, and the spatial scale of local adaptation is likely to be smaller, making a smaller smoothing radius appropriate. In cases where gene flow is relatively high, genotypes are likely to be mixed over a larger spatial neighborhood, making a larger smoothing radius appropriate. (In cases where drift is relatively high, the phenotypic character of populations can diverge in ways that may be adaptively neutral or potentially maladaptive.) Since the true patterns of local adaptation are never known, comparing various choices for this parameter can help to assess this source of uncertainty.

2.1.7 *Climate versus soil* Climate and soil conditions both influence plant fitness, but their relative importance varies by species. By default, the five climate variables and five soil variables that underlie the tool are given equal weight in calculating environmental differences and seed zones, but this can be changed by adjusting the "soil versus climate" slider. This can be useful if species-specific knowledge suggests that

9

one or the other category is particularly important for a given species, or as a way of evaluating how sensitive the results are to assumptions about this topic. See the data and methods section for more details about the climate and soil variables used in the model.

2.2 Interpreting results

Any time one of the above input settings is changed, a new set of results is calculated and displayed on the map and scatter plot. The scatter plot includes one point for every grid cell in the species range map (except for species with very large ranges, in which case a random sample of several thousand points is shown.) The color of grid cells and points matches, and represents the output variable selected in the "display variable" menu. The variables represented on the x and y dimensions of the scatter plot can also be changed using the respective drop-down menus.

2.2.1 *Environmental difference* This is a measure of how different a location's environment is at one point in time from the focal site's environment at another point in time, based on the set of soil and climate variables used in the tool. It is the default display variable represented by color, and is the only display variable that is specific to the selected focal site. It is a multidimensional z-score, representing the equivalent number of one-dimensional standard deviations between the focal site environment and the location environment. The standard deviation is defined based on historic spatial variation across the species range. The greater this value, the more dissimilar a location's environment is from the focal site's environment.
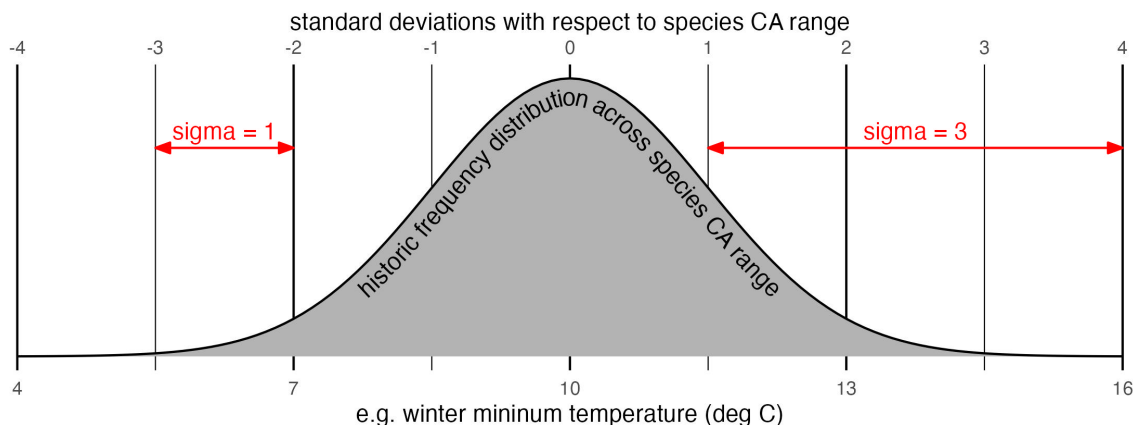

Figure 1: **Illustration of the metric used to calculate multidimensional environmental differences. The difference between two sites' environments is expressed as a sigma (σ) value that is equivalent to a one-dimensional standard deviation.**

2.2.2 *Seed zones* This feature helps to visualize patterns of environmental variation across a species' range by clustering sites with similar climate and soil characteristics into discrete groups or zones within a species range. The slider that appears when "seed zones" is selected as the display variable can be used to select the desired number of clusters. The seed zones are sensitive to the "soil versus climate" slider and to the "smoothing radius" slider, but they do not depend on the location of the selected focal site. See the technical methods section for more details on how seed zones are calculated.

2.2.3 ***Climate variables*** The tool is based on 5 climate variables that are broadly relevant to the distribution and function of plants in California. These can be used as scatter plot axes as well as color variables, helping users to explore patterns of environmental variation across species ranges. The climate data used here are from CHELSA (*3*). Spatially smoothed historic climate values are displayed when focal site activity is set to "planting," while non-smoothed projected future climate for the selected time period and emissions scenario are displayed when it is set to "collection."

2.2.4 ***Soil variables*** Soil similarity is based on 16 chemical and physical soil variables relevant to plants at different soil depths, reduced to 5 uncorrelated dimensions via a statewide principal component analysis. These principal components (PC) can be used as scatter plot dimensions and as color variables. See the methods section for further detail about the particular soil properties associated with each PC. The soil data used here are from SoilGrids (*4*). Smoothed values are displayed when focal site activity is set to "planting," while non-smoothed values are displayed when it is set to "collection."

2.2.5 ***Climate change variables*** For each climate variable, users can also select the projected amount of local climate change as a display variable. This represents the mean projected local climate change for an ensemble of 5 GCMs for the selected time period and emissions scenario, relative to the baseline time period. The app also includes a variable called "multivariate change in climate," which uses the same statistical metric as the "environmental difference" variable described above, but ignores soil variables and measures local change at each site relative to its own baseline climate (rather than relative to a shared focal site). No spatial smoothing is applied to these climate change variables.

3 METHODS AND DATA

3.1 Climate

Seeds of Change is based on five climate variables broadly relevant to California plant ecology:

- Winter minimum temperature: average nighttime minimum temperature from December–February.
- Summer maximum temperature: average daytime high temperature from June–August.
- Total annual precipitation: total precipitation, log-transformed for ecological relevance and statistical normality
- Annual climatic water deficit (CWD): difference between potential and actual evapotranspiration, a measure of aridity, integrated over monthly temperature and precipitation
- Annual actual evapotranspiration (AET): a measure of water and energy simultaneously available for plant growth

As noted above, the tool uses data from four time periods (a historic period and three future periods) and three future emissions scenarios (SSPs). The data represent the average climate over 30-year periods at a 1 km spatial scale. Historical data are from CHELSA version 2.1 (*3*). Monthly precipitation and minimum, mean, and maximum temperature data for the baseline time period, as well as for each future time period (n = 3), CMIP6 emissions scenario (n = 3), and general circulation model (GCM) (n = 5), were used to derive the five climate variables employed in the app, with CWD and AET computed using the 'hydro' R package (*5*). Results are based on ensemble means across the 5 GCMs for a given year and scenario, though data for individual models are shown for the focal site to give a sense for uncertainty.

3.2 Soils

Soil data are from SoilGrids (*4*), and encompass 16 chemical (e.g. pH, cation exchange capacity, and carbon content) and physical soil variables (e.g. sand, silt, clay, and coarse fragment content) at various soil depths. Data from from SoilGrids version 1 were used in lieu of the more recent version 2, because the latter omits developed areas of California that are often important for restoration projects. Soil data for 16 variables, each at 4 depths were downloaded at 250 m resolution, aggregated to the 1 km CHELSA grid by averaging, and masked to the State of California. Figure 2 lists the variables and depths. Values for each variable were then standardized and normalized using a Yeo-Johnson power transformation (*6*). The variable set was then reduced using a principal component analysis (PCA), retaining the first 5 principal components, which collectively capture more than 90% of the variation in the original set of soil variables. Figure 2 shows the spatial patterns in each of these variables, and their associations with the input variables.

3.3 Species distributions

The species distribution models used in this app are from Kling *et al.* (*7*). Maxent was used to model suitability based on herbarium specimen collection locations and four climate variables (annual climatic water deficit, annual precipitation, summer maximum temperature, and winter minimum temperature), and these suitabilities were then down-weighted in sites with high human impact and in sites geographically far from known species occurrences. More details can be found in the publications documenting the original
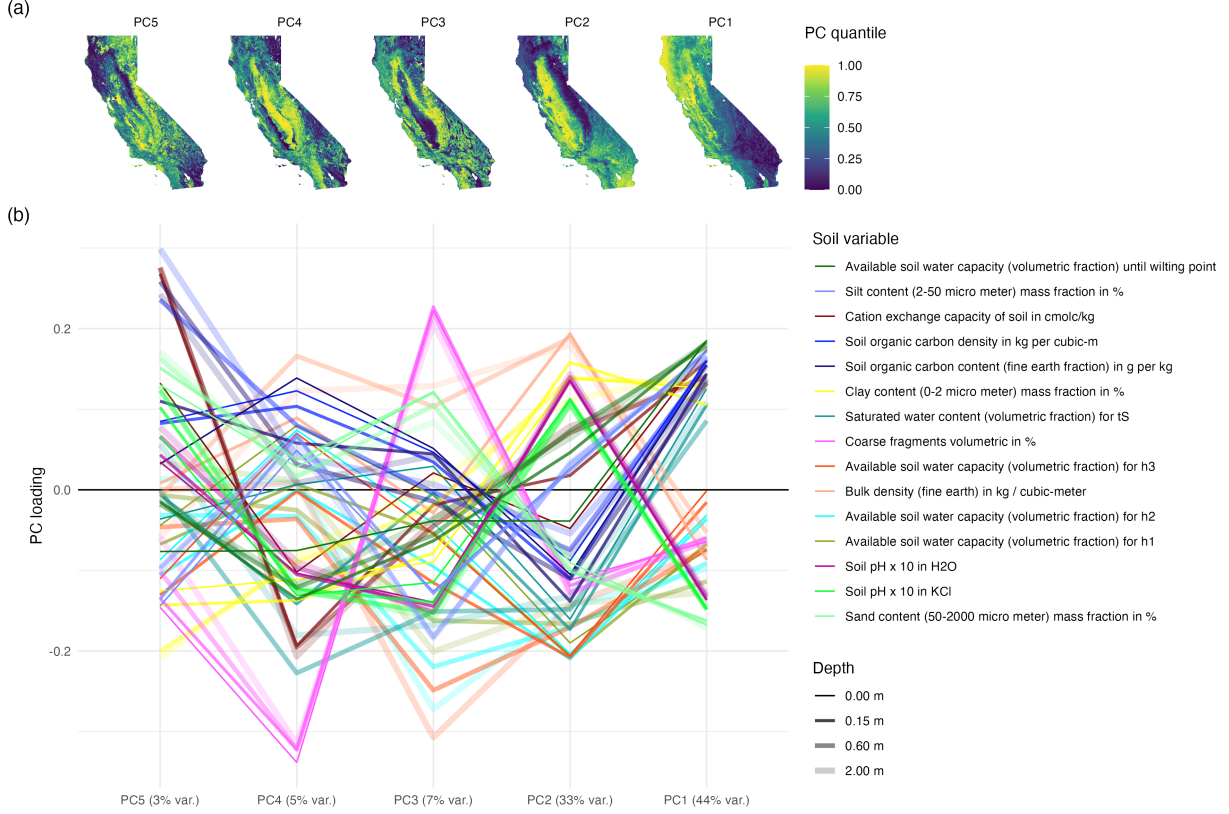
Figure 2: Principal components (PCs) of soil variation across California. (a) Maps of the first 5 PCs, each rank-transformed for visual clarity. (b) Relationship between these PCs and the PCA input variables, with larger positive (negative) values indicating a stronger positive (negative) variable loading.

models (*7, 8*).

The *Seeds of Change* tool thresholds these continuous occurrence estimates to generate binary range maps for further analysis. As described above, users can set the threshold value using the "SDM threhshold" slider. These ranges are only model estimates and will contain a mix of false positives and false negatives.

3.4 Environmental difference

The environmental difference between two sites, σ, is calculated as

$$\sigma = \sqrt{qchisq(pchisq(\sum_{i=1}^{k}(kp_i(f_i - x_i)^2), k), 1)}$$

where $qchisq$ and $pchisq$ are the quantile and distribution functions of the Chi-squared distributions, respectively, k is the number of environmental variables, f and x are length-k vectors of univariate z-scores for the focal site and the evaluation site, respectively, and p is a length-k vector of variable weights summing to 1. The functions pchisq and qchisq are the distribution function and quantile function of the chi-squared

distribution, with k and 1 degrees of freedom, respectively. The z-scores f and x are standardized with respect to the mean and standard deviation of each environmental variable across the species range during the historic period. The variable p is determined by the user-selected "soil vs climate" input, with all soil variables getting one weight and climate variables getting a second. This formula is modified from Mahony et al. (*9*).

This equation effectively expresses the difference between multivariate z-scores in terms of the difference between univariate z-scores, in order to aid interpretation. Many users will be familiar with univariate z-scores, in which the units are standard deviations of the reference distribution. When $k = 1$, the above equation reduces to the much simpler $\sigma = f - x$. For higher dimensionalities—i.e. in our case when k = 10— this formulation maintains the property that if the focal site's environment were equal to the multivariate mean environment across the species range, 68%, 95%, and 99.7% of sites across the range would have σ values less than 1, 2, and 3, respectively, during the baseline time period.

Seed zones

Seed zones are estimated by using a k-means clustering algorithm to group the sites across a species range into a user-specified number of clusters. In order to accommodate user-specified soil versus climate weighting, prior to clustering, environmental variables are standardized across a species range, and then scaled by the square root of p, where p is a weighting variable as described above.

K-means clustering is performed using the Hartigan-Wong algorithm (*10*). This is a non-deterministic method that involves some internal randomization; a consistent seed is used for the random number generator in the *Seeds of Change* app to ensure reproducibility, but it should be noted that the app returns but one among an array of possible clustering solutions.

3.5   Source code

All code for the *Seeds of Change* app and input data processing is open source, and available at `https://github.com/matthewkling/seedsource`. Please report bugs by opening a GitHub issue in the repository at that link, or by emailing mattkling@berkeley.edu.

REFERENCES

(1) Richardson, D. M. et al. (2009). Multidimensional evaluation of managed relocation. *Proceedings of the National Academy of Sciences 106*, 9721–9724.

(2) Schwartz, M. W. et al. (2012). Managed relocation: Integrating the scientific, regulatory, and ethical challenges. *BioScience 62*, 732–743.

(3) Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data 4*, 1–20.

(4) Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One 12*, e0169748.

(5) Kling, M. hydro: Compute Water Balance Variables, R package version 0.1.0, 2023.

(6) Weisberg, S. (2001). Yeo-Johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June 1*, 2003.

(7) Kling, M. M., Mishler, B. D., Thornhill, A. H., Baldwin, B. G., and Ackerly, D. D. (2019). Facets of phylodiversity: evolutionary diversification, divergence and survival as conservation targets. *Philosophical Transactions of the Royal Society B 374*, 20170397.

(8) Thornhill, A. H., Baldwin, B. G., Freyman, W. A., Nosratinia, S., Kling, M. M., Morueta-Holme, N., Madsen, T. P., Ackerly, D. D., and Mishler, B. D. (2017). Spatial phylogenetics of the native California flora. *BMC biology 15*, 1–18.

(9) Mahony, C. R., Cannon, A. J., Wang, T., and Aitken, S. N. (2017). A closer look at novel climates: new methods and insights at continental to landscape scales. *Global Change Biology 23*, 3934–3955.

(10) Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*, 100–108.