Received: 23 March 2022 | Accepted: 11 May 2022

DOI: 10.1002/ece3.8981

Ecology and Evolution Open Access WILEY

RESEARCH ARTICLE

A comparison in species distribution model performance of succulents using key species and subsets of environmental predictors

Catherine E. Buckland[1] | Andrew J. A. C. Smith[2] | David S. G. Thomas[1,3]

[1]School of Geography and the Environment, University of Oxford, Oxford, UK

[2]Department of Plant Sciences, University of Oxford, Oxford, UK

[3]Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa

Correspondence
Catherine E. Buckland, School of Geography and the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK.
Email: catherine.buckland@ouce.ox.ac.uk

Funding information
Oxford Martin School, Grant/Award Number: Dryland Bioenergy

Abstract

Identifying the environmental drivers of the global distribution of succulent plants using the Crassulacean acid metabolism pathway of photosynthesis has previously been investigated through ensemble-modeling of species delimiting the realized niche of the natural succulent biome. An alternative approach, which may provide further insight into the fundamental niche of succulent plants in the absence of dispersal limitation, is to model the distribution of selected species that are globally widespread and have become naturalized far beyond their native habitats. This could be of interest, for example, in defining areas that may be suitable for cultivation of alternative crops resilient to future climate change. We therefore explored the performance of climate-only species distribution models (SDMs) in predicting the drivers and distribution of two widespread CAM plants, *Opuntia ficus-indica* and *Euphorbia tirucalli*. Using two different algorithms and five predictor sets, we created distribution models for these exemplar species and produced an updated map of global inter-annual rainfall predictability. No single predictor set produced markedly more accurate models, with the basic bioclim-only predictor set marginally out-performing combinations with additional predictors. Minimum temperature of the coldest month was the single most important variable in determining spatial distribution, but additional predictors such as precipitation and inter-annual precipitation variability were also important in explaining the differences in spatial predictions between SDMs. When compared against previous projections, an *a posteriori* approach correctly does not predict distributions in areas of ecophysiological tolerance yet known absence (e.g., due to biotic competition). An updated map of inter-annual rainfall predictability has successfully identified regions known to be depauperate in succulent plants. High model performance metrics suggest that the majority of potentially suitable regions for these species are predicted by these models with a limited number of climate predictors, and there is no benefit in expanding model complexity and increasing the potential for overfitting.

KEYWORDS
Crassulacean acid metabolism (CAM), *Euphorbia tirucalli*, Hellmann–Eberle quotient, *Opuntia ficus-indica*, species distribution model (SDM), succulents

TAXONOMY CLASSIFICATION
Agroecology; Applied ecology; Biogeography; Ecophysiology

1 | INTRODUCTION

Identifying the environmental conditions under which a species can thrive is an important question in biogeography and ecology both to understand the environmental tolerances of individual organisms and to be able to predict their distributions across current and future climates. Many parts of the world are likely to experience warmer climates and reduced and/or more variable precipitation in the decades ahead, so there is interest in determining which organisms may be relatively well adapted to these future climate regimes. A group of plants that are particularly characteristic of warm, semi-arid parts of the world with strong seasonal rainfall patterns are succulents using the specific mode of photosynthesis known as Crassulacean acid metabolism (CAM). By virtue of being able to fix most of their carbon dioxide from the atmosphere at night rather than during the day time, CAM plants typically show high water-use efficiency and can survive in environments with high daily temperatures and relatively limited water availability (Cushman, 2001; Lüttge, 2010; Osmond, 1978; Winter, 1985; Winter & Smith, 1996). The environmental resilience of these plants makes them attractive species for cultivation on marginal land for a variety of potential uses, such as fodder, bioethanol production, or as feedstock for anaerobic digestion (Acharya et al., 2019; Borland et al., 2009; Davis et al., 2011; Hastilestari et al., 2013; Holtum et al., 2011; Loke et al., 2011; Mason et al., 2015; Mwine et al., 2013; Yan et al., 2011). Such crops may be of particular value in semi-arid regions most likely to experience increased drought risk (e.g., Marthews et al., 2019; Otto et al., 2018).

The growth and the ecophysiological controls on the natural distribution of CAM species have been widely studied and observed across a range of environments. Broadly speaking, the methods previously used to observe the distribution of specific CAM species can be split into those that are: observation based; growth/trial based; and those that are based on models—both process and data-driven (Ringelberg et al., 2020). However, a comparison of the importance in different environmental parameters and derived indices in explaining the variability in CAM plant distribution has not yet been completed. Using existing studies published in the literature it is possible to compare areas of expected growth and productivity suitability (i.e., the locations with the environmental conditions required for specific species growth) (Guisan et al., 2017) based on process-based models (e.g., Owen et al., 2015) or using climatic envelope methods (e.g., Louhaichi et al., 2015). However, there is also the potential to use methods based on derived environmental parameters and those driven by *a posteriori* models (e.g., species distribution modeling (Guisan et al., 2017)) to identify the relationship between known observations of CAM species and predictor variables; thus projecting maps of suitable biotic conditions for species to occur based on climatological, environmental, and/or biotic correlations (Aguirre-Gutiérrez et al., 2013; Soberón & Nakamura, 2009).

Correlative species distribution models (SDMs) have been commonly employed as predictive tools to quantify relationships between species occurrence datasets and measurements of environmental variables (Dormann et al., 2012) across ecology, but seldom applied to the specific mapping of CAM plants. Equally, as noted by Bucklin et al. (2015), there remains no consensus on which variables should be included as predictors in SDM analysis more generally. While many climate-only SDMs (i.e., using only climatic parameters) have been highlighted as important tools for both projecting current and future ecological niches (e.g., for guiding future conservation efforts (Elith & Leathwick, 2009) (Bucklin et al., 2015)), some studies have criticized this approach for providing only an incomplete representation of complex environmental systems (Araújo & Peterson, 2012; Bahn & McGill, 2007; Beale et al., 2008; Heikkinen et al., 2006). Using different combinations of bioclimatic and derived environmental indices, this study tests and compares the relative importance of parameters in explaining the distribution of CAM plants, focusing specifically on *Opuntia ficus-indica* (L.) Mill. And *Euphorbia tirucalli* L. as example species. In doing so, this study attempts to define the best, minimal predictors of plant distribution so that models have the greatest predictive power without being over-parameterized (Merow et al., 2014; Raes & Aguirre-Gutiérrez, 2018).

Unlike recent analyses which have ensemble-modeled numerous species with the aim of identifying the wider natural succulent biome distribution (e.g., Ringelberg et al., 2020), this study takes an alternative approach by selecting a minimal number of species of interest, but for which their distribution is successfully wide, occupying all available climatic niches, and with minimal dispersion limitations. There are numerous rare succulent species that have very restricted ranges on account of being dispersal-limited for which this analysis would not be appropriate. By comparison, *O. ficus-indica* is a successful invasive species having established itself across every continent (except Antarctica) (CABI, 2019) and found across all latitudes. *Opuntia ficus-indica* and *E. tirucalli* have also shown great potential suitability for bioeconomic uses (Hastilestari et al., 2013; Mason et al., 2015); and are therefore suitable test species to use for this analysis which is interested in exploring the possibility for these plants to be actively grown as a crop—highlighting the potential that can be achieved with CAM plantation for bioeconomic and land restorative purposes. While most previous distribution modeling exercises have been built on the natural distribution of native species, additional novel information might be obtained from explicitly considering the extent

TABLE 1 Bioclim parameters (from Fick & Hijmans, 2017) used in the final model iterations

Bioclim variable	Environmental parameter
Bioclim 2	Mean diurnal temperature range (mean of monthly (max temp-min temp)) (°C)
Bioclim 6	Minimum temperature of coldest month (°C)
Bioclim 12	Annual precipitation (mm)
Bioclim 15	Precipitation seasonality (coefficient of variation)

and spread of introduced invasive species, once they are given the opportunity to spread into other parts of the "potential niche."

Specifically, this study will compare different sets of variables to predict zones of potential suitability for *Opuntia ficus-indica* and *Euphorbia tirucalli* growth. In doing so, this study aims to first predict the current locations with suitable biotic conditions for the occurrences of *O. ficus-indica* and *E. tirucalli* using different SDMs tested in this study. Second, the results will help identify the most important set of variables that help define the environmental niche of two CAM species of interest. While the natural distribution of both species has generally been restricted to semi-arid regions as outcompeted by other plants, their natural ecological requirements permit them growing in wetter areas, and competition factors have largely restricted the spread of the species to regions with annual rainfall <500 mm (Luttge, 2004). *Opuntia ficus-indica* is a successful invasive which has been widely sighted across regions outside of central America (e.g., Africa, southern Europe), while *E. tirucalli* is native to Africa (Palgrave, 1977; Webb et al., 1984) but has also been found in central America, Europe, and other locations globally. Given the successful expansion, but different origins of these two species, comparison of the potential regions through which they could be successfully cultivated for bioeconomic (e.g., biogas) uses across a region (e.g., sub-Saharan Africa) with low levels of energy access, increased agricultural pressure in the face of drought, and high climatic suitability for these species is particularly interesting (Buckland & Thomas, 2021). For this reason, this study will initially calibrate and project models based on a global view, before taking a deeper focus on Africa as a potential region for cultivation, bioenergy and bioeconomic uses.

2 | MATERIALS AND METHODS

Using SDM techniques, this study compares the relative performance of five SDMs to predict the potential distribution of *O. ficus-indica* and *E. tirucalli* based on current climatic conditions. The five SDMs each capture different combinations of environmental variables defined in the WorldClim 2.1 bioclim database (Fick & Hijmans, 2017) and derived indices or parameters that have previously been cited as impacting upon the spatial distribution of CAM plants: the Hellmann–Eberle quotient (a measure of inter-annual rainfall predictability used by Ellenberg, 1981), the aridity index (the ratio between annual precipitation and potential evapotranspiration (PET)), cloud cover (as a proxy for light intensity), and the R-index (the ratio between actual and PET) (Yao, 1974). As noted in Title and Bemmels (2018), the inclusion of more complex climatic indices may characterize environmental conditions that are more directly physiologically relevant to particular species than more primary climatic parameters (e.g., temperature, precipitation). Due to the successful invasive nature of both species, we have considered their expansion to be largely limited by environmental conditions rather than distribution-limited, and thus only climatic-based parameters have been used.

2.1 | Predictor datasets

The choice of environmental variables selected should ideally be based on the known ecology of the species (Title & Bemmels, 2018), as this has previously demonstrated more realistic SDMs (Rödder et al., 2009; Saupe et al., 2012). With this in mind, a combination of bioclim datasets from the WorldClim 2.1 catalogue (Fick & Hijmans, 2017) and derived environmental metrics were compiled and a sensitivity analysis (Pearson's Correlation) was used to remove highly correlated variables. Inclusion of co-variant parameters leads to over-parameterization of the model. All predictor datasets were bilinearly resampled to the same 2.5 min resolution.

2.1.1 | Bioclim datasets

Based on existing research of the parameters impacting the growth and distribution of succulents and CAM plants more generally (Acharya et al., 2019; Inglese & Scalenge, 2009; Le Houérou, 1996; Louhaichi et al., 2015; Masocha & Dube, 2018), and the results from covariance testing (Appendix A), four bioclim variables were selected for use as explanatory parameters (Table 1).

2.1.2 | Hellmann–Eberle quotient

The Hellmann–Eberle quotient provides a measure of inter-annual precipitation variability and is defined as the ratio between precipitation of the wettest year and precipitation of the driest year over an extended period of time. Ellenberg (1981) examined the distribution pattern of tall stem succulents in relation to climate and found that they tended to occur in areas where rainfall was low (i.e., <500 mm per annum), but regularly received (i.e., where the Hellmann–Eberle quotient <5 over a series of years) (Cowling et al., 1997). Ellenberg's original study from 1981 was based on 35 years of observations (1905–1940) and has since been referred to and expanded in more recent studies exploring the controls on CAM distribution (e.g., Holtum et al., 2016, 2017; Lüttge, 2010; Ringelberg et al., 2020). Using historical monthly weather data from 1960 to 2018 AD from the CRU-TS 4.03 dataset (Harris et al., 2014) downscaled with WorldClim 2.1 (Fick & Hijmans, 2017), we calculated a more recent version of the Hellmann–Eberle quotient based on annual historical precipitation levels at a 2.5 min spatial resolution (globally) to compare against observational occurrences of *O. ficus-indica* and *E. tirucalli* from the Global Biodiversity Information Facility (GBIF.org, 2020). Individual GeoTiff files were analyzed and climate rasters were produced in R Studio (RStudio Team, 2019), before being combined with observational occurrence data in ArcGIS Pro 2.4.1.

Precipitation regime alone, however, is unlikely to explain the distribution of these species as it does not include the impact of minimum temperatures, which is known to be limiting for particular CAM species (Acharya et al., 2019; Herrando-Moraira, 2020; Inglese & Scalenge, 2009; Smith et al., 2012; Stock et al., 1997). For this reason, combining the Hellmann–Eberle quotient with other bioclimatic parameters in the SDM analysis has the potential to improve our distributional understanding of key species of interest.

2.1.3 | Aridity index, R-index and cloud cover

The Aridity Index (AI) is commonly considered to provide a measure of overall water availability, a central component to all vegetative growth. Based on global raster data from 1970 to 2000 AD, a global aridity index based upon the implementation of the Penman–Monteith reference evapotranspiration equation (Allen et al., 1998) was used in this study (Trabucco & Zomer, 2018). The R-index is calculated as the ratio between actual evapotranspiration (AET) and PET and is a measure of plant water supply in relation to plant water demand (Yao, 1974). A global R-index raster was calculated using the average annual AET and PET rates available via the Consultative Group for International Agricultural Research (Trabucco & Zomer, 2018). Finally, as a proxy for photosynthetically active radiation, cloud cover was included as a potential parameter that could be inversely related to plant growth. CAM plant growth shows a saturation-type relationship to light intensity (Nobel, 1988; Nobel & Valenzuela, 1987) with the three main environmental limitations on CAM plant growth considered water, light, and temperature (Nobel, 1988; Owen et al., 2015). Process-based models have thus included a proxy for light intensity as a measure to predict the variability in spatial productivity of CAM plant species in existing literature (e.g., Owen et al., 2015). In this study, a global raster of mean annual cloud cover based on 15 years (2000–2014 AD) of twice-daily satellite observations was used from the EarthEnv data repository (Wilson & Jetz, 2016).

2.1.4 | Pearson's correlation coefficient

A total of five combinations of environmental parameters and bioclim parameters (Fick & Hijmans, 2017) were used to model the relationship between environmental conditions and the observed distribution of *O. ficus-indica* and *E. tirucalli* (Table 2). Prior to final environmental parameter selection for each of the five SDM combinations, Pearson's correlation coefficient tests were conducted to test for covariance between the variables (Appendix A). Based on the results, and on an understanding of the main climatic parameters that influence CAM distribution, 4 bioclim variables were selected for use in the final model fitting (Table 1) alongside a combination of derived environmental indices.

2.2 | Occurrence data

Opuntia ficus-indica and *Euphorbia tirucalli* were the two species of interest selected for analysis in this study. The former is an especially suitable test species for this analysis since its occurrences are already occupying most of its geographic range allowing us to model a potential distribution closer to its fundamental niche (i.e., all the environmental conditions where a species could potentially exist) as opposed to the realized niche (i.e., those conditions in which the species currently does exist) (Chase & Leibold, 2003; Hutchinson, 1957). By comparison, often the current distributions of localized or very rare species are restricted by dispersal limitations and species interactions; in such cases the realized niche will be smaller than the fundamental niche, and we cannot independently test the impact of different climatic and environmental parameters on defining areas suitable for species occurrence.

Opuntia ficus-indica and *E. tirucalli* occurrence data were downloaded from the GBIF data repository (GBIF.org, 2020) (Accessed 09/06/2020) and cleaned according to the method described in Zizka (2019). Species occurrence data from both the native and introduced ranges was used for both species. One of the main aims of this study is to identify regions which could support the cultivation of these species under current climatic conditions (i.e., to map the fundamental niche of the species). As such, we do not need to limit the training dataset to the native distribution, rather observations of the species across a range of geographic zones are useful in identifying the scope of environmental settings which are suitable. Spatial bias of occurrence datasets has the potential to distort the interpretation of large-scale biodiversity patterns (Ballesteros-Mejia et al., 2013; Beck et al., 2014; Boakes et al., 2010; Varela et al., 2014; Yang et al., 2013), and SDMs are sensitive to the spatial bias of specimen records (Dudík & Phillips, 2005; Lintz et al., 2013; Phillips et al., 2009). Spatially biased data would have a two-fold impact on distorting SDMs: first, through biasing the present data used to train and evaluate model performance (Hijmans et al., 2017); second in biasing the surface range envelope model used in the pseudo-absence dataset generation (see below) and therefore model performance metrics. With this in mind, we applied a geographic sampling filter, selecting up to five occurrence data points from each 1° × 1° grid cell—reducing our datasets to 2721 and 1085 occurrences (from 8061 and 2313) of *O. ficus-indica* and *E. tirucalli*, respectively (Figures 1 and 2).

2.3 | Pseudo-absences

Unlike "presence" datasets, "absence" datasets are not often readily available. Since some SDM algorithms require both datasets, pseudo-absence (PA) datasets are created as a replacement for true absence records (Raes & Aguirre-Gutiérrez, 2018). The use of PA data is widely accepted and has been shown in the SDM literature to be a useful approach to calibrate SDMs (Chefaoui & Lobo, 2008; Iturbide et al., 2018; Václavík & Meentemeyer, 2009; Wisz & Guisan, 2009). PA data are generated by sampling background areas from which presence records have not been identified through a range of different strategies, including: random, surface range envelope (SRE), or based on a minimum (or maximum) distance from known presence points. The sensitivity of SDM algorithms to the sample of PA when projecting under future climates varies between models and creates a source of SDM-dependent uncertainty that should be considered when deciding on initial PA sampling and accounted for in SDM ensemble modeling (Iturbide et al., 2018).

Based on the recommendations of the findings in Barbet-Massin et al. (2012) and Iturbide et al. (2018), an equal number of PAs were selected to presences with multiple PA realizations (five) to reduce overall uncertainty. Studies based on a single realization of PAs have the potential to mask results from poorly performing SDMs (Iturbide et al., 2018). Hence, five PA realizations were used to reduce the dependence on poorly performing SDMs and to ensure model fits were not dependent on a single realization where PAs have been biasedly generated from regions with few noted presences rather than few true presences (Barbet-Massin et al., 2018). PAs were generated from all areas outside the suitable area estimated by a surface range envelope model (SRE) (Thuiller et al., 2014). SRE models are based on presence-only data (Barbet-Massin et al., 2012); SRE quantile refers to the quantile used to remove the most extreme values of each environmental variable for determining tolerance boundaries (quantile 0.025 ~ 95% confidence interval) (Hallgren et al., 2019).

2.4 | Model fitting

There are numerous options for algorithms to use in SDM studies (summarized in Raes & Aguirre-Gutiérrez, 2018), but there is often no model of "best" choice (Qiao et al., 2015). Fitting the data with the same algorithm over multiple repeats would yield different results, as would fitting the data across multiple algorithms. Overfitting occurs when an overly flexible model learns the noise in the training dataset to a level that negatively impacts the performance of the model when introduced to new input data. By comparison, inflexible models do not have the flexibility to fit complex relationships between parameters and predictor datasets. As such, inflexible models may not have the capacity to accurately fit the training dataset, nor to generalize well to new unseen data (e.g., projecting over a new time period or geographic location). In SDM, and machine learning more generally, we seek to find a balance in creating models with the capacity to fit variance but also avoid bias. Equally, defining "best" model is largely dependent on the choice of evaluative

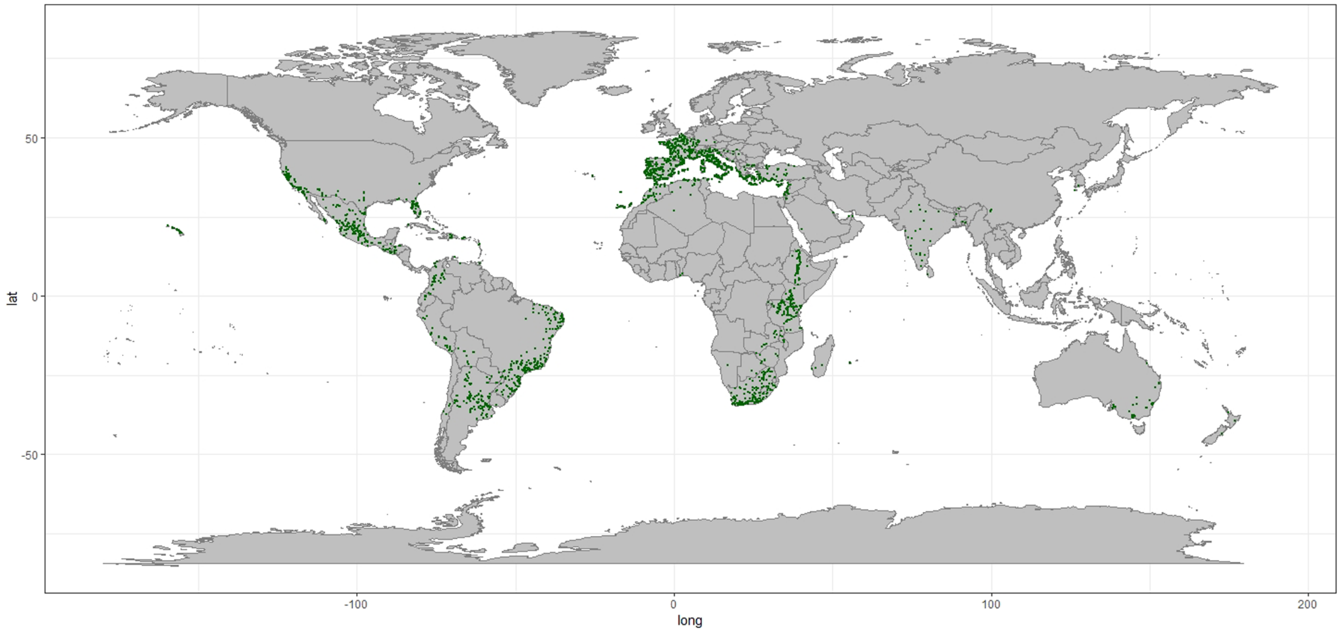

FIGURE 1 Final 2721 *Opuntia ficus-indica* occurrences downloaded from the GBIF dataset (GBIF.org, 2020) after spatial bias analysis completed

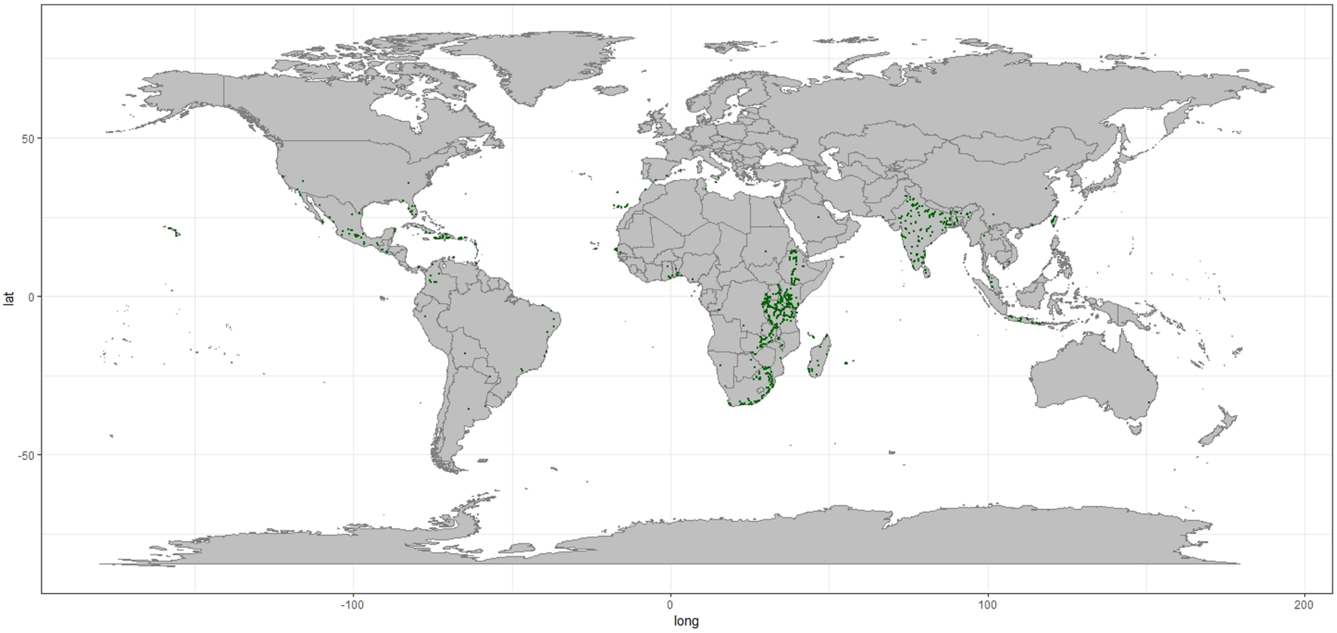

FIGURE 2 Final 1085 *Euphorbia tirucalli* occurrences downloaded from the GBIF dataset (GBIF.org, 2020) after spatial bias analysis completed

metric—for which there are numerous. Each evaluative metric measures a slightly different aspect of model performance, and thus while a model may perform well according to one measure, it may not be the "best" model according to another metric (Qiao et al., 2015; Raes & Aguirre-Gutiérrez, 2018).

With this in mind, SDMs were initially fitted across two different algorithms which required the same PA dataset generation strategy: Boosted Regression Trees (Elith, 2008) (also known as Generalized Boosting Model GBM) and Random Forests (Breiman, 2001), before being combined in an ensemble model to obtain a consensus distribution (Marmion et al., 2009). Default model parameters found in the biomod2 package (Georges & Thuiller, 2013; Thuiller et al., 2014) were used and 10 repeats were completed per algorithm per PA selection, producing a total of 100 individual model repeats used for each ensemble model (a total of 500 individual model repeats across all five SDM scenarios). The between- and within-modeling variability shown in SDM outputs has led to the widespread usage of ensemble models (Marmion et al., 2009; Qin et al., 2020; Raes & Aguirre-Gutiérrez, 2018; Senay et al., 2013); capturing the uncertainty in model predictions across the different SDM algorithm outputs (Araújo & New, 2007; Dormann, 2018; Hao et al., 2019; Raes & Aguirre-Gutiérrez, 2018), producing more consistent predictions when projecting new unseen data (e.g., future climate scenarios).

There are many strategies that can be used to combine predictions from individual models into an ensemble model. Following the recommendation of Hao et al. (2019), we have taken a more sophisticated approach which involved weighting the models based on their individual predictive performances. The performance of each individually trained model was assessed, and ensemble models were produced based on the true skill statistic (TSS) and relative operating characteristic (ROC) of each individual model (based on thresholds defined in Qin et al. (2020): those with ROC >0.5 imply that the model performed better than random). TSS metrics are widely used as a measure of relative performance in SDM studies and have been recommended over the use of other methods such as Kappa (Allouche et al., 2006). The TSS is calculated as: *Specificity +Sensitivity –1*, whereby "specificity" refers to the proportion of correctly predicted absences, and "sensitivity" refers to the proportion of correctly predicted presences. Individual models were combined using two ensemble-model algorithms: weighted mean of probabilities and coefficient of variation of probabilities, to provide a measure of uncertainty in the former ensemble model. Current occurrence and predictor datasets were split 60% for training and validation, with the remaining 40% used for testing and evaluating model performance. All models were fitted and projected using the biomod2 package version 3.3 (Thuiller et al., 2014) in R Studio version 1.2.5033 (RStudio Team, 2019).

2.5 | Evaluating model comparison

As well as TSS and AUC (ROC) scores calculated for each of the individual models, the TSS and AUC scores of the ensemble models were compared to determine the relative best performing model and identify whether the additional parameters used in SDMs 2–5 increased the predictive accuracy of SDM 1 (bioclim-only predictors). As discussed in Komac et al. (2016), the AUC provides us with a measure of the performance of ordinal score models and a threshold measure of accuracy (Thuiller et al., 2005), while the TSS score provides us with a measure of evaluative performance which has all the advantages associated with the Cohen's kappa statistic (Cohen, 1968) but is not sensitive to prevalence (Allouche et al., 2006). Ensemble models from the five SDM scenarios were initially projected on to the world to generate a continuous map showing variations in the suitability/probability of occurrence for the two species of interest. Then, using the ensemble model cut-off values to provide a binary measure of habitat suitability, projections were then compared against projections based on existing methods from the literature (e.g., Louhaichi et al., 2015) to identify the spatial variability in identified suitable regions between the methods. Ensemble binary cut-off values are calculated as those that give the maximum "sensitivity" and "specificity" scores (Thuiller et al., 2005).

2.6 | Assessing variable importance

Individual variable importance was approximated using the Variables_importance function of the "biomod2" package (Thuiller et al., 2014). Variable importance was assessed for each of the five ensemble models and across each of the two species with the aim of determining which climatic or environmental factors have stronger effects on the species suitability across the region of interest. The principle of the biomod2 variable importance algorithm is to shuffle a single variable of the given data and produce model predictions with this new "shuffled" dataset. A Pearson's correlation between the reference predictions and "shuffled" dataset predictions is calculated, with higher values corresponding to a greater influence the individual variable has on the model (i.e., a value of 0 assumes no influence of the variable on the model). Variable importance results were standardized across all predictors used per model and presented in percentage terms.

3 | RESULTS

3.1 | Ensemble model projections and comparisons

A total of 500 individual models and projections were produced for each species and ensembled to produce weighted mean projections with coefficient of variation (uncertainty between the individual projections) measurements for each of the scenarios. Ensemble results from SDM scenario 1 are presented in Figures 3 and 4, with TSS scores across the five ensembles shown in Table 3. SDM scenarios 2–5 are shown in Appendix E.

Across both species, results show relatively little difference in the evaluative performance between the ensemble models when tested against the remaining 40% of the dataset, however, SDM 1 outperformed the other four SDMs for both *O. ficus-indica* and *E. tirucalli* distribution projections (Table 3, Figures S7 and S8). The random forest algorithm generally performed best for *O. ficus-indica* projections in both TSS and ROC scores, while GBMs marginally outperformed random forest models in the *E. tircualli* predictions (See Supplementary Information). Among both species and predictor scenarios, all models performed well with overall TSS scores >0.91 across all ensembles (Table 3). TSS scores for individual model

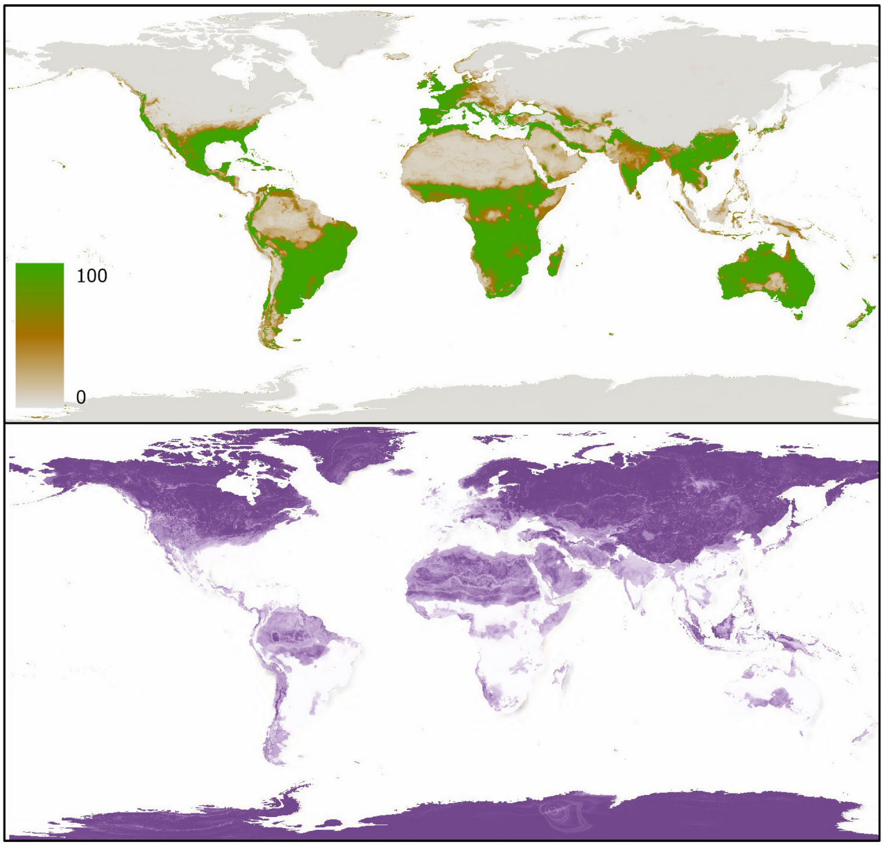

FIGURE 3 Species distribution model scenario 1 projection and uncertainty (coefficient of variation) based on occurrences of *O. ficus-indica*

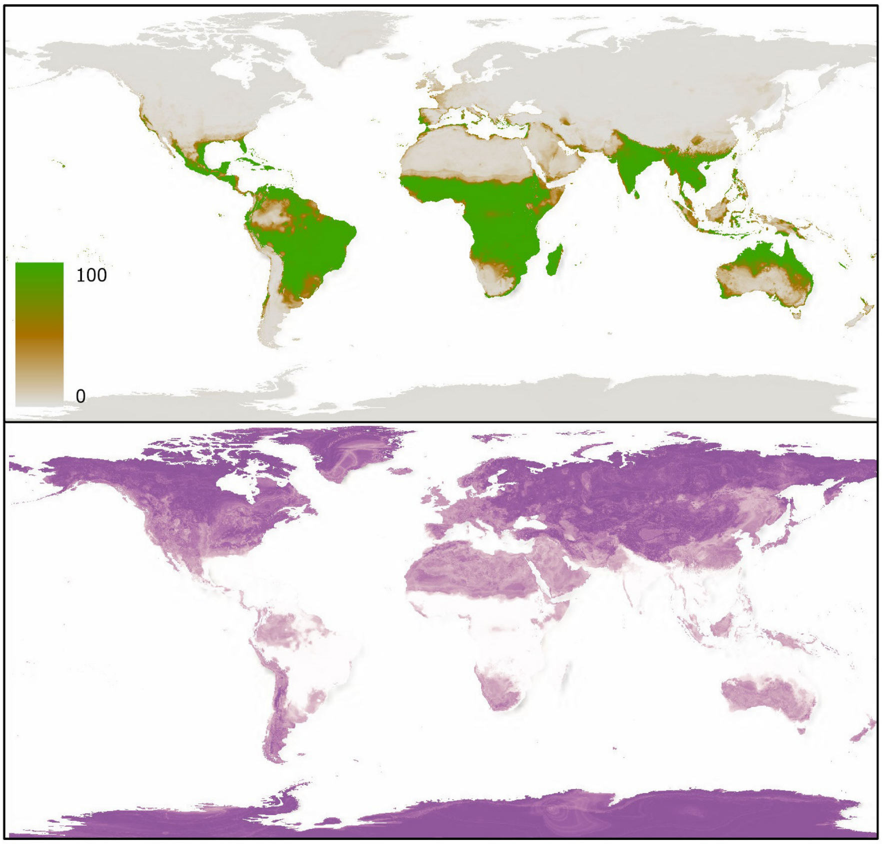

FIGURE 4 Species distribution model scenario 1 projection and uncertainty (coefficient of variation) based on occurrences of *E. tirucalli*

TABLE 3 Ensemble species distribution model (SDM) evaluative metrics (true skill statistic (TSS) and relative operating characteristic (ROC)) for each of the five *O. ficus-indica* and *E. tirucalli* ensembled SDM scenarios. See Supplementary Information for binary cut-off, Specificity, and Sensitivity scores

	O. ficus indica		*E. tirucalli*	
SDM scenario	TSS	ROC	TSS	ROC
1	0.930	0.997	0.955	0.998
2	0.914	0.994	0.932	0.996
3	0.916	0.995	0.948	0.997
4	0.925	0.996	0.954	0.998
5	0.918	0.995	0.949	0.997

performances showed high performance with little variability, ranging from 0.85 to 0.942 and from 0.87 to 0.968 for *O. ficus-indica* and *E. tirucalli*, and 0.981–0.996 and 0.984–0.999 ROC scores, respectively (Tables S2 and S3). Due to the overall high performance of the individual models, all individual projections were included in the weighted ensemble model.

At a global scale, ensemble models across all five predictor scenarios indicated that both species have potential distributional ranges in the tropics and mid-latitudes. The areas predicted most suitable for *O. ficus-indica* include sub-Saharan Africa, Mediterranean Europe, Australia, South America (especially Brazil and north-eastern Argentina), central America and countries in southern and eastern Asia (e.g., India, China, and Thailand). Meanwhile the areas predicted most suitable for *E. tirucalli* growth are more restricted to the tropics, especially sub-Saharan Africa, Brazil and northern South America, India, northern Australia and south China. The higher latitudes and hyper-arid Sahara were predicted unsuitable for both species.

When the deviation in environmental suitability is compared between SDM scenarios (Figures 5 and 6), the inclusion of either the Hellmann–Eberle quotient, aridity index, or R-index all produced overall results with lower suitability projections than those predicted using bioclim variables alone (SDM 1). It is only in SDM 4 (Figures 5c and 6c) that ensemble model projections suggest that some regions (typically those with reduced overall certainty) have a higher level of environmental suitability than projections based on the four bioclim variables alone. However, these results are not necessarily corroborated when we consider the binary cutoff values at a regional scale for example (i.e., where maximum specificity and sensitivity are achieved) and the results are presented as either “suitable” or “unsuitable” areas (Table 4). For example, results from the continuous profiles suggest SDM 4 estimates some areas of both increased and decreased suitability relative to SDM 1, yet the results from the binary cutoff values for the African continent suggest this projection produces the second lowest levels of regions suitable for *O. ficus-indica* growth. By comparison, SDM 4 produces the largest suitable area estimates for the *E. tirucalli* projections, as well as demonstrating increased estimated suitability values in the continuous dataset for SDM 4 relative to SDM 1.

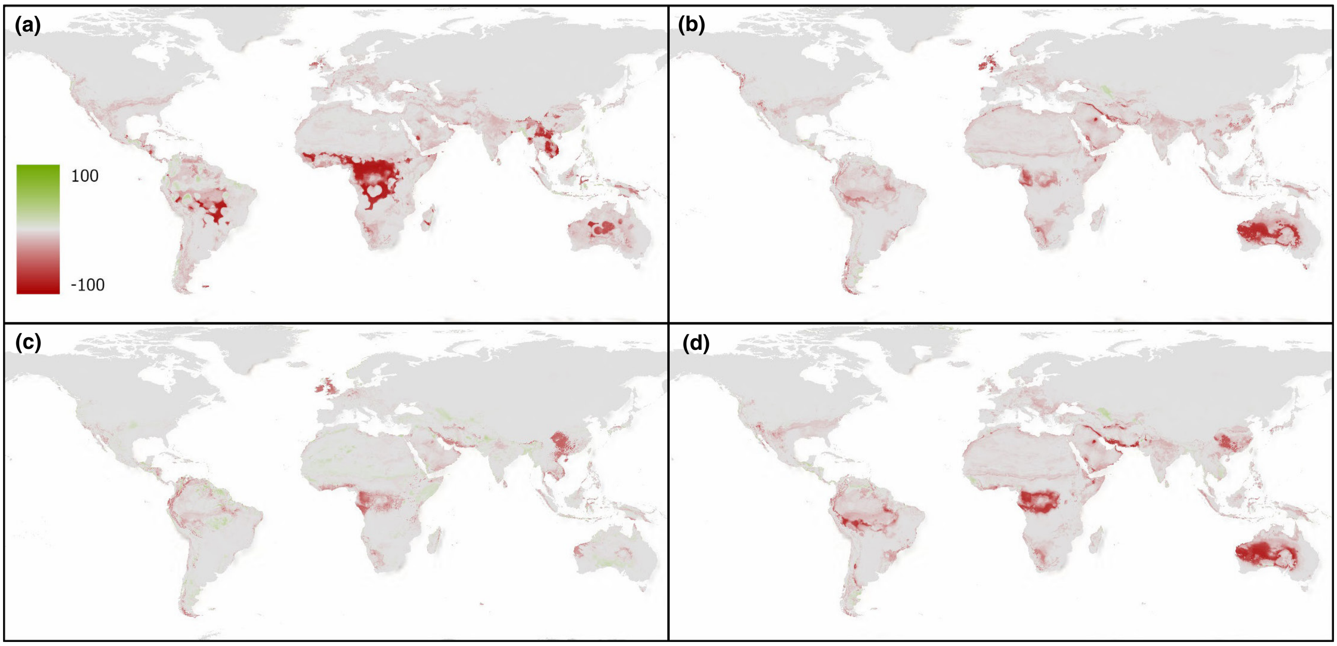

FIGURE 5 Deviation of species distribution model (SDM) scenarios 2–5 (a–d) from the results of the bioclim-only scenario (SDM 1) for *O. ficus-indica*. Red shading indicates areas where the relative SDM predicts a lower probability of *O. ficus-indica* growth versus SDM 1, while green shading predicts areas with a higher probability of *O. ficus-indica* projected occurrence

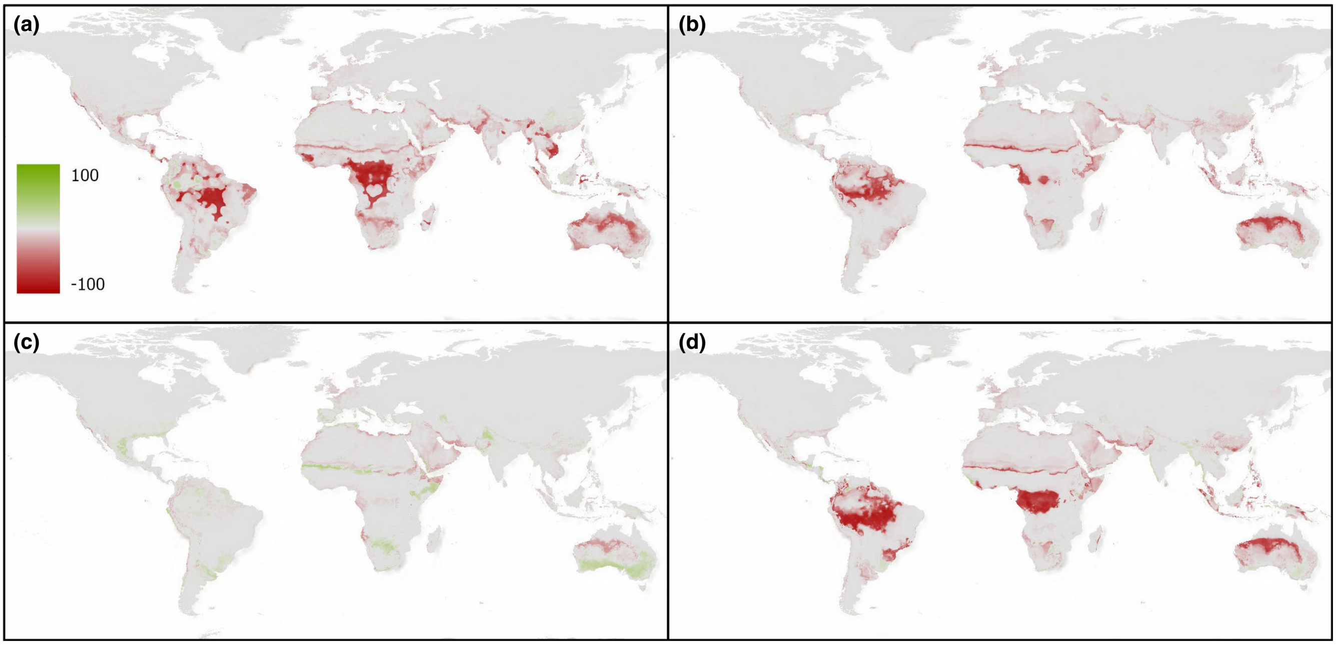

FIGURE 6 Deviation of species distribution model (SDM) scenarios 2–5 (a–d) from the results of the bioclim-only scenario (SDM 1) for *E. tirucalli*. Red shading indicates areas where the relative SDM predicts a lower probability of *E. tirucalli* growth versus SDM 1, while green shading predicts areas with a higher probability of *E. tirucalli* projected occurrence

3.2 | Environmental variable importance

Results from individual variable importance analysis were calculated based on the weighted mean ensemble models for each of the five SDM scenarios and across the two species of interest (Tables 5 and 6). Across both *O. ficus-indica* and *E. tirucalli*, the minimum temperature of the coldest month shows a significantly higher variable importance factor than any of the other environmental parameters across the SDM scenarios. Equally, both species show similarity in response to annual precipitation, which demonstrates second greatest individual variable importance, except for when modeled in scenarios 3 and 5—when the Aridity index and R-index, respectively, show high levels of variable importance and a reduction in the relative importance of annual precipitation.

4 | DISCUSSION

4.1 | Drivers of CAM plant distribution

Results from the ensemble model evaluative performance and individual variable importance analysis suggest that for both species there is not any overall model improvement with the inclusion of either the aridity index, Hellmann–Eberle quotient, cloud cover conditions or R-index (i.e., SDMs 2–5) over the primary four bioclim variables (SDM1); and that the dominant variable of importance in explaining the spatial variability in ecological niche is the minimum temperature of the coldest month. With this in mind, it seems there is little benefit in the inclusion of additional predictors beyond the four bioclim parameters, regardless of which additional parameters were to be considered. With results not differing significantly between the SDM scenario analyses, it suggests that the most important bioclimate predictors (SDM 1) primarily shaped the patterns across all models produced. These results of variable importance are in agreement with von Willert et al. (1992) who consider low temperatures a key limiting factor in succulent growth when referring to succulent growth on hill slopes in Tenerife. The relatively minor variation in overall model performance between the SDMs with and without the additional parameters is also in agreement with the results noted by Bucklin et al. (2015), who have suggested that climate-only predictor sets may be equally as effective in producing environmental suitability maps.

TABLE 4 Example total suitable area (million km^2) calculations across the African continent (as an example) for *O. ficus-indica* and *E. tirucalli* per species distribution model (SDM) scenario based on the binary cutoff values

SDM scenario	*O. ficus-indica*	*E. tirucalli*
1	15.6	17.0
2	11.4	13.2
3	14.8	16.3
4	14.4	17.4
5	13.4	14.9

Following the role of extreme cold temperatures, moisture availability measured either through annual precipitation or the aridity index or R-index is shown to be the second most important independent variable on overall model performance. When an alternative precipitation metric is included in the model (i.e., SDM scenarios 3 and 5), the relative importance of annual precipitation is reduced. The compound variable, aridity index, is defined as the ratio between annual precipitation and PET—reflecting the amount of moisture potentially available for vegetation growth. Equally, the R-index as calculated as the ratio between AET and PET, provides a measure of water supply in relation to water demand (Yao, 1974); unsurprising that the relative importance of annual precipitation as an individual metric is reduced when considered alongside these compound variables. However, it is also worth noting that the R-index used in this study (derived from AET and PET datasets (Trabucco & Zomer, 2018)) is based on spatially standardized vegetation and soil coefficients (i.e., based on typical

TABLE 5 Standardized mean variable importance of each parameter across the five different species distribution model (SDM) scenarios for *O. ficus-indica*

	Environmental predictors							
SDM scenario	Mean diurnal temp range	Min temp of coldest month	Annual precipitation	Precipitation seasonality	Hellmann–Eberle quotient	Aridity index	Cloud cover	R-index
1	2%	74%	20%	4%	n/a	n/a	n/a	n/a
2	1%	70%	17%	4%	8%	n/a	n/a	n/a
3	2%	74%	5%	4%	n/a	15%	n/a	n/a
4	1%	78%	13%	5%	n/a	n/a	4%	n/a
5	1%	75%	5%	4%	n/a	n/a	n/a	14%

TABLE 6 Standardized mean variable importance of each parameter across the five different species distribution model (SDM) scenarios for *E. tirucalli*

	Environmental predictors							
SDM scenario	Mean diurnal temp range	Min temp of coldest month	Annual precipitation	Precipitation seasonality	Hellmann Eberle quotient	Aridity index	Cloud cover	R-index
1	2%	71%	26%	1%	n/a	n/a	n/a	n/a
2	2%	68%	23%	1%	7%	n/a	n/a	n/a
3	2%	74%	3%	1%	n/a	19%	n/a	n/a
4	1%	82%	14%	1%	n/a	n/a	2%	n/a
5	2%	76%	4%	1%	n/a	n/a	n/a	18%

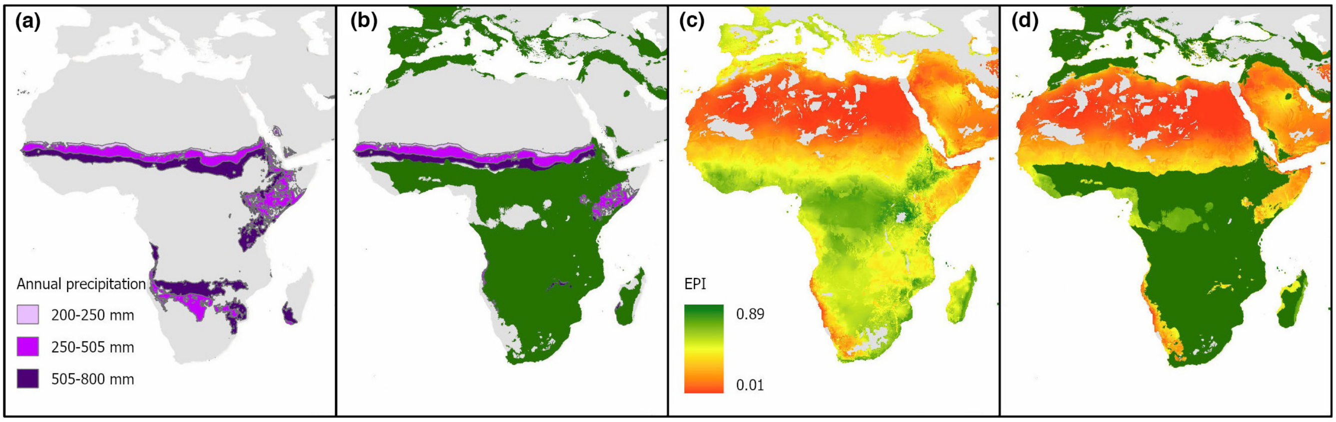

FIGURE 7 Comparison of species distribution model (SDM) 1 binary *O. ficus-indica* projected ecological niche with existing methods from the literature: (a–b) estimates of potential area suitable for *O. ficus-indica* growth based on the method described in Louhaichi et al. (2015) overlain with SDM 1 binary projections (this study) (c - d) refined Environmental Productivity Index (EPI) for *O. ficus-indica* as calculated in Owen et al. (2015) overlain with SDM 1 binary projections (this study)

agronomic crops at maturity and an average soil texture for plant rooting depth at 2 m). Variations in both the vegetation and soil stress coefficients specific to the characteristics of the species of interest would perhaps produce a more useful spatial representation and metric to test.

Moreover, it is important to note that the variable importance results reported refer to the individual direct influence of that variable on the model projection, it does not account for *interactions* between the variables or combined effects of the variables—a key tenet of SDM approaches. For example, while cloud cover has in general shown low levels of individual variable importance, Figure 6c demonstrated that SDM scenario 4 was the only ensemble projection to identify an increase in land suitability estimates from the bioclim-only model—suggesting that the role of cloud cover (or rather the inverse) is significant in determining the ecological niche of *E. tirucalli*, albeit likely through interactions with other variables. Equally, despite the consistently high TSS values and lack of variability between the different SDM predictor sets studied (Table 3), the spatial distribution in the ecological niche suitability estimates is shown to vary between scenarios (Figures 5 and 6). These results suggest that despite marginal variation in TSS score or variable importance factors, the interactions between variables are important in explaining the overall projected suitability profile for individual species, and the minimum temperature of the coldest month, while important, is not exclusively the sole variable which defines the distribution of either species. Rather, it is the combination of both parameters documenting minimum temperatures, and also a measure of precipitation (both in terms of annual total amount, and/or a measure of variability in precipitation) which are important in explaining the ecophysiological controls on these species.

This being said, while the results in the spatial deviation of individual SDM scenarios from SDM 1 projections (Figures 5 and 6) suggest variation in the continuous likelihood profiles, binary cut-off levels (Table 4) suggest that all alternative (i.e., SDMs 2–5) SDMs for *O. ficus indica* predict a reduction in suitable area relative to SDM 1, while *E. tirucalli* results suggest SDM 4 projects marginally greater levels of suitability than SDM 1 when assessed at a continent-scale, for example. Thus, while the continuous suitability profiles may show one measure of difference between the alternative predictor scenarios, the binary levels of "suitable" versus "unsuitable" areas in absolute terms provide an alternative interpretation of the overall size of the ecological niche. Nevertheless, despite these values suggesting >15 million km^2 of suitable area for *O. ficus indica* (e.g., SDM 1) across Africa, the potential yields will vary within these locations/SDM projections and hence a combination of both the continuous scale likelihood and the binary cut-off values is useful in assessing the true scale of potential niche that could be used for growing these species.

4.2 | Land suitability estimates

A key advantage of the SDM approach is the capacity to produce a more refined estimate of land area that is potentially available, after taking account of protected areas and other essential land covers and uses, for cultivation of *O. ficus-indica* and *E. tirucalli*. Given the overall lack of variability found between the five SDM scenarios and the equally high performance of the bioclim-only SDM 1 model, the following section opted to only compare the results from the *O. ficus-indica* SDM 1 model with existing methods from previous literature focusing specifically on Africa as an example region. Figure 7 presents the comparison of the land suitability estimates found in this study following SDM 1 (binary cut-off) and the predicted suitable areas for *O. ficus-indica* growth according to the parameters detailed in Louhaichi et al. (2015), and the adapted productivity index displayed in Owen et al. (2015).

Figure 7b overlays the results from this study onto the theoretical distribution of *O. ficus-indica* across Africa according to the parameters detailed in Louhaichi et al. (2015) (Figure 7a),

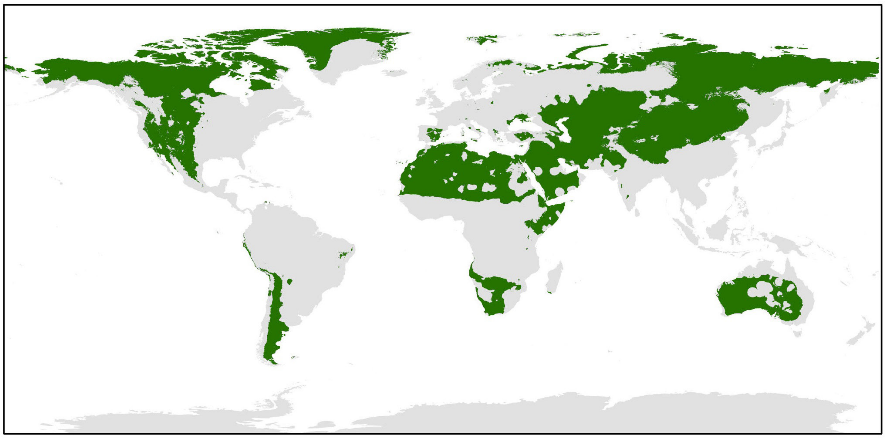

FIGURE 8 Based on the monthly precipitation values from 1960 to 2018, average annual precipitation, the Hellmann–Eberle quotient (maximum annual precipitation/minimum annual precipitation), and overall Ellenberg index were calculated. Green areas represent regions where <500 mm of rainfall coincides with Hellmann–Eberle quotient >5. Global raster of Ellenberg index as shown was used as a predictor dataset for species distribution model 2

with results showing additional theoretically suitable areas in northern Africa bordering the Mediterranean, a greater region in eastern Africa, and more extensive suitability in southern Africa. SDM 1 projected distribution details a far greater suitable area than the approach taken in Louhaichi et al. (2015) since they are based on observed occurrence data rather than restricted by the common intersection of a few environmental conditions. While the models used in this study do not consider any soil-based parameters, they have still explained over 93% of the occurrences observed with high AUC scores. When compared with the results of the productivity analysis in Owen et al. (2015), our results show a clear omission of *O. ficus-indica* growth in central Africa where Figure 7c suggests a zone of high productivity. This is a good demonstration that our approach has taken the "competition" aspect into consideration as the EPI method suggests that *O. ficus-indica* would grow well in central Africa, but we know through lack of occurrences in these areas that *O. ficus-indica* is out-competed by other plants.

Unlike the two alternative methods described above, the SDM method explored in this study is driven by the relationship with known occurrences and climatic parameters, allowing us to qualify these maps with a level of evaluative performance. As noted earlier, this suggests that c.1,500 million hectares of land are suitable for *O. ficus-indica* and *E. tirucalli* growth and is of importance to initiatives looking at the potential use of CAM plant biomass as feedstock for anaerobic digestion and bioenergy, or alternative hydrolysis and VFA uses such as bioplastics, proteins. The advantage of an SDM-based approach which incorporates the nuances and complexities of the relationships between environmental parameters and known occurrences, is that while tropical areas are theoretically identified of potential high productivity, *O. ficus-indica* is outcompeted and occurrence data demonstrates that it is not a successful plant in these regions for reasons beyond its direct relationship with climate. This conclusion is key to identifying the most appropriate regions for exploring the potential for cultivation of CAM plants, such as *O. ficus-indica* and *E. tirucalli*, as it removes any discussion regarding the removal of prime forest ecosystems in place of CAM cultivation.

4.3 | Updated Hellmann–Eberle quotient map

While minimum temperatures were demonstrated as key in determining the majority of the variability in spatial distribution of the species, analysis of an updated Ellenberg index (Hellmann–Eberle quotient combined with average annual precipitation) also highlighted the importance of precipitation predictability in the distribution of succulents. As noted above, Ellenberg (1981) examined the distribution pattern of tall stem succulents in relation to climate (von Willert et al., 1992) and found that they tended to occur in areas where rainfall was low (i.e., <500 mm), but regularly received (Hellmann–Eberle quotient <5 over a long series of years) (Cowling et al., 1997). Since Ellenberg's original study, which was based on precipitation data from 1905 to 1940, further studies have also explored the predictability of rainfall as a parameter by which to explain succulent distributions (Holtum et al., 2016; Ringelberg et al., 2020). As part of this study, an updated global Hellmann–Eberle quotient based on a longer time-series of monthly precipitation data from 1960 to 2018 was used as a predictor parameter for the ensemble model. In addition to use in the ensemble modeling, the updated map of a revised "Ellenberg index" shown in Figure 8 provides further valuable discussion to unresolved problems regarding succulent distribution. The near absence of stem succulents from arid Australia, for example, is one particular example which has invited discussion among research groups (Holtum et al., 2016; Ringelberg et al., 2020). While Ellenberg (1981) suggested the rainfall is too unpredictable to support stem succulents in arid Australia, Ringelberg et al.'s (2020) recent ensemble model of the wider succulent biome has suggested that large parts of Australia should be climatically suitable for stem succulents; further complicating their apparent absence. Instead, Ringelberg et al. (2020) suggest that perhaps longer-term climatic oscillations, or even historical fire conditions, may offer an alternative rationale for their absence despite favorable climatic conditions, according to their ensemble models.

By comparison, the updated Hellmann–Eberle quotient and "Ellenberg index" maps produced using a much longer period of climate data (58 years) in this study have successfully identified regions that are well-known areas depauperate in succulents, like central

Australia and large parts of Kalahari/Namib deserts. Additionally, it is highlighting other areas that agree well with observation—parts of the Arabian Peninsula, Horn of Africa, Saharan desert, and in South America the Atacama. This updated visualization based on a longer time series than previously studied suggests that perhaps high variability in annual precipitation levels over the long term is key to explaining succulent absence, such as the lack of endemic terrestrial species with CAM in arid Australia.

5 | CONCLUSIONS

In comparison with existing methods of land suitability estimation for these species, this study has taken an a posteriori modeling approach using SDMs and known occurrences to extrapolate wider areas of potential suitability for cultivation of these species. In doing so, it has allowed us to qualify the models of suitability estimates with a level of evaluative performance, incorporates the nuances and complexities of relationships between environmental parameters and known occurrences, and produce a more refined estimate of land area that is potentially available for cultivation of *O. ficus-indica* and *E. tirucalli* when considered alongside existing land uses and primary ecosystems. The high model performance metrics of SDMs made using successfully invasive distribution-unlimited species gives us confidence that most of the fundamental niche of *O. ficus-indica* and *E. tirucalli* can be explained by the models produced in this study, and given the negligible variability between the different scenarios, there is no benefit in expanding model complexity and increasing the potential for over-fitting by including additional environmental predictors. While the minimum temperature of the coldest month was found to be the key variable of importance in determining the spatial variability of *O. ficus-indica* and *E. tirucalli*, these results are based on the individual performance of each parameter as opposed to combined effects and nonlinearities between the environmental predictors. An updated global map of Hellmann–Eberle quotient based on a much longer period of climate data (ca. 60 years), supports the ideas of Ellenberg (1981) that long-term precipitation variability is also a key variable in determining CAM plant distribution, and in certain regions can explain stem succulent absence.

AUTHOR CONTRIBUTION

Catherine E. Buckland: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Writing – original draft (lead); Writing – review & editing (lead). **Andrew J.A.C. Smith:** Formal analysis (supporting); Funding acquisition (equal); Investigation (supporting); Supervision (supporting); Writing – review & editing (supporting). **David S. G. Thomas:** Funding acquisition (equal); Investigation (supporting); Supervision (supporting).

ACKNOWLEDGEMENTS

This research was funded by the Oxford Martin School as part of the Dryland Bioenergy project. We would also like to extend our gratitude to the two anonymous reviewers for providing valuable feedback on an earlier draft of this manuscript.

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The authors confirm that the results supporting the findings of this study are available within the article and its supplementary materials. Raw data and code (R script) to reproduce the results are available at Dryad: https://doi.org/10.5061/dryad.wwpzgmsmt.

ORCID

Catherine E. Buckland https://orcid.org/0000-0002-7411-3046
Andrew J. A. C. Smith https://orcid.org/0000-0001-9188-0258
David S. G. Thomas https://orcid.org/0000-0001-6867-5504

REFERENCES

Acharya, P., Biradar, C., Louhaichi, M., Ghosh, S., Hassan, S., Moyo, H., & Sarker, A. (2019). Finding a suitable niche for cultivating cactus pear (Opuntia ficus-indica) as an integrated crop in resilient dryland agroecosystems of India. *Sustainability (Switzerland)*, *11*, https://doi.org/10.3390/su11215897

Aguirre-Gutiérrez, J., Carvalheiro, L. G., Polce, C., van Loon, E. E., Raes, N., Reemer, M., & Biesmeijer, J. C. (2013). Fit-for-purpose: Species distribution model performance depends on evaluation criteria - dutch hoverflies as a case study. *PLoS One*, *8*(5), e63708. https://doi.org/10.1371/journal.pone.0063708

Allen, R. G., Periera, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration: guideline for computing crop water requirement. In FAO Irrigation and Drainage, Paper No 56. FAO, Rome, Italy; 300.

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, *22*, 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Araújo, M. B., & Peterson, A. T. (2012). Uses and misuses of bioclimatic envelope modelling. *Ecology*, *93*, 1527–1539.

Bahn, V., & McGill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*, 733–742. https://doi.org/10.1111/j.1466-8238.2007.00331.x

Ballesteros-Mejia, L., Kitching, I. J., Jetz, W., Nagel, P., & Beck, J. (2013). Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography*, *22*, 586–595. https://doi.org/10.1111/geb.12039

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barbet-Massin, M., Rome, Q., Villemant, C., & Courchamp, F. (2018). Can species distribution models really predict the expansion of invasive species? *PLoS One*, *13*, 1–14. https://doi.org/10.1371/journal.pone.0193085

Beale, C. M., Lennon, J. J., & Gimona, A. (2008). Opening the climate envelop reveals macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences USA*, *105*, 14908–14912.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLOS Biology*, *8*, e1000385. https://doi.org/10.1371/journal.pbio.1000385

Borland, A. M., Griffiths, H., Hartwell, J., & Smith, J. A. C. (2009). Exploiting the potential of plants with crassulacean acid metabolism for bioenergy production on marginal lands. *Journal of Experimental Botany*, *60*, 2879–2896. https://doi.org/10.1093/jxb/erp118

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Buckland, C. E., & Thomas, D. S. G. (2021). Analysing the potential for CAM-fed bio-economic uses in sub-Saharan Africa. *Applied Geography*, *132*, 102463. https://doi.org/10.1016/j.apgeog.2021.102463

Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., Speroterra, C., & Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, *21*, 23–35. https://doi.org/10.1111/ddi.12247

CABI (2019). Invasive Species Compendium - Opuntia ficus-indica (prickly pear). Centre for Agriculture and Bioscience International. Accessed October 15, 2019. https://www.cabi.org/isc/datasheet/37714

Chase, J. M., & Leibold, M. A. (2003). *Ecological niches: Linking classical and contemporary approaches*. University of Chicago Press.

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, *210*, 478–486. https://doi.org/10.1016/j.ecolmodel.2007.08.010

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. https://doi.org/10.1037/h0026256

Cowling, R. M., Richardson, D. M., & Pierce, S. M. (1997). *Vegetation of southern Africa*. Cambridge University Press.

Cushman, J. C. (2001). Crassulacean acid metabolism. A plastic photosynthetic adaptation to arid environment. *Plant Physiology*, *127*, 1439–1448.

Davis, S. C., Dohleman, F. G., & Long, S. P. (2011). The global potential for Agave as a biofuel feedstock. *GCB Bioenergy*, *3*, 68–78. https://doi.org/10.1111/j.1757-1707.2010.01077.x

Dormann, C. F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*, 485–504.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., & Singer, A. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography*, *39*, 2119–2131. https://doi.org/10.1111/j.1365-2699.2011.02659.x

Dudík, M., & Phillips, S. J. (2005). Correcting samle selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, *18*, 323–330.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

Ellenberg, H. (1981). Reasons for stem succulents being present or absent in the arid regions of the world. *Flora*, *171*, 114–169.

Fick, S., & Hijmans, R. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*, 4302–4315. https://doi.org/10.1002/joc.5086

GBIF.org. (2020). GBIF Occurrence Download. https://doi.org/10.15468.dl.9fff6p

Georges, D., & Thuiller, W. (2013). Multi-species Distribution Modeling with biomod2, 1–11.

Guisan, A., Thuiller, W., & Zimmermann, N. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., & Mackey, B. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, *408*, 108719. https://doi.org/10.1016/j.ecolmodel.2019.108719

Hao, T., Elith, J., Guillera-Arroita, G., & Lahoz-Monfort, J. J. (2019). A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions*, *25*, 839–852. https://doi.org/10.1111/ddi.12892

Harris, I., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations - the CR TS3.10 Dataset. *International Journal of Climatology*, *34*, 623–642.

Hastilestari, B. R., Mudersbach, M., Tomala, F., Vogt, H., Biskupek-Korell, B., Van Damme, P., Guretzki, S., & Papenbrock, J. (2013). Euphorbia tirucalli L.-comprehensive characterization of a drought tolerant plant with a potential as biofuel source. *PLoS One*, *8*, 1–12. https://doi.org/10.1371/journal.pone.0063501

Heikkinen, R. K., Luoto, M., Araújo, M. B., Virkkala, R., Thuiller, W., & Sykes, M. T. (2006). Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, *30*, 751–777. https://doi.org/10.1177/0309133306071957

Herrando-Moraira, S., Vitales, D., Nualart, N., Gómez-Bellver, C., Ibáñez, N., Massó, S., Cachón-Ferrero, P., González-Gutiérrez, P. A., Guillot, D., Herrera, I., Shaw, D., Stinca, A., Wang, Z., & López-Pujol, J. (2020). Global distribution patterns and niche modelling of the invasive Kalanchoe × houghtonii (Crassulaceae). *Scientific Reports*, *10*, 3143. https://doi.org/10.1038/s41598-020-60079-2

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). Dismo R Package (version 1.1-4). https://cran.r-project.org/package=dismo

Holtum, J. A. M., Chambers, D., Morgan, T., & Tan, D. K. Y. (2011). Agave as a biofuel feedstock in Australia. *GCB Bioenergy*, *3*, 58–67. https://doi.org/10.1111/j.1757-1707.2010.01083.x

Holtum, J. A. M., Hancock, L. P., Edwards, E. J., Crisp, M. D., Crayn, D. M., Sage, R., & Winter, K. (2016). Australia lacks stem succulents but is it depauperate in plants with crassulacean acid metabolism (CAM)? *Current Opinion in Plant Biology*, *31*, 109–117. https://doi.org/10.1016/j.pbi.2016.03.018

Holtum, J. A. M., Hancock, L. P., Edwards, E. J., & Winter, K. (2017). Facultative CAM photosynthesis (crassulacean acid metabolism) in four species of Calandrinia, ephemeral succulents of arid Australia. *Photosynthesis Research*, *134*, 17–25. https://doi.org/10.1007/s11120-017-0359-x

Hutchinson, G. E. (1957). Concluding remarks. *Population Studies: Animal Ecolgy and Demography*, 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

Inglese, P., & Scalenge, R. (2009). Cactus pear (Opuntia ficus-indica L. (Mill)). In E. A. C. Constantini (Ed.), *Manual of methods for soil and land evaluation* (pp. 275–285). Science Publisher.

Iturbide, M., Bedia, J., & Gutiérrez, J. M. (2018). Background sampling and transferability of species distribution model ensembles under climate change. *Global and Planetary Change*, *166*, 19–29. https://doi.org/10.1016/j.gloplacha.2018.03.008

Komac, B., Esteban, P., Trapero, L., & Caritg, R. (2016). Modelization of the current and future habitat suitability of rhododendron ferrugineum using potential snow accumulation. *PLoS One*, *11*, 1–18. https://doi.org/10.1371/journal.pone.0147324

Le Houérou, H. N. (1996). The role of cacti (Opuntia spp.) in erosion control, land reclamation, rehabilitation and agricultural development in the Mediterranean Basin. *Journal of Arid Environments*, *33*, 135–159. https://doi.org/10.1006/jare.1996.0053

Lintz, H. E., Gray, A. N., & McCune, B. (2013). Effect of inventory method on niche models: Random versus systematic error. *Ecological Informatics*, *18*, 20–34. https://doi.org/10.1016/j.ecoinf.2013.05.001

Loke, J., Mesa, L. A., & Franken, Y. J. (2011). Euphorbia tirucalli bioenergy manual: Feedstock production, bioenergy conversion, applications, economics version 2.

Louhaichi, M., Park, A. G., Mata-Gonzalez, R., Johnson, D. E., & Mohawesh, Y. M. (2015). A Preliminary Model of Opuntia ficus-indica (L.) Mill. Suitability for Jordan A Preliminary Model of Opuntia ficus-indica (L.) Mill. Suitability for. *Acta Horticulturae*, *1067*, 267–274. https://doi.org/10.17660/ActaHortic.2015.1067.37

Luttge, U. (2004). Ecophysiology of crassulacean acid metabolism. *Annals of Botany*, *93*, 629–652.

Lüttge, U. (2010). Ability of crassulacean acid metabolism plants to overcome interacting stresses in tropical environments. *AoB PLANTS*, *2010*, 1–15. https://doi.org/10.1093/aobpla/plq005

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, *15*, 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x

Marthews, T. R., Jones, R. G., Dadson, S. J., Otto, F. E. L., Mitchell, D., Guillod, B. P., & Allen, M. R. (2019). The impact of human-induced climate change on regional drought in the horn of Africa. *Journal of Geophysical Research: Atmospheres*, *124*, 4549–4566. https://doi.org/10.1029/2018JD030085

Masocha, M., & Dube, T. (2018). Global terrestrial biomes at risk of cacti invasion identified for four species using consensual modelling. *Journal of Arid Environments*, *156*, 77–86. https://doi.org/10.1016/j.jaridenv.2018.05.006

Mason, P. M., Glover, K., Smith, J. A. C., Willis, K. J., Woods, J., & Thompson, I. P. (2015). The potential of CAM crops as a globally significant bioenergy resource: Moving from "fuel or food" to "fuel and more food". *Energy and Environmental Science*, *8*, 2320–2329. https://doi.org/10.1039/c5ee00242g

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., Mcmahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*, 1267–1281. https://doi.org/10.1111/ecog.00845

Mwine, J., Van Damme, P., Hastilestari, B. R., & Papenbrock, J. (2013). African natural plant products: Discoveries and challenges in chemistry, health and nutrition. *American Chemical Society*, *2*, 4905–4914.

Nobel, P. S. (1988). *Environmental biology of agaves and cacti*. Cambridge University Press.

Nobel, P. S., & Valenzuela, A. G. (1987). Environmental responses and productivity of the CAM plant, Agave tequiliana. *Agricultural and Forest Meteorology*, *39*, 319–334.

Osmond, C. B. (1978). Crassulacean acid metabolism: A curiosity in context. *Annual Review of Plant Physiology*, *29*, 379–414. https://doi.org/10.1146/annurev.pp.29.060178.002115

Otto, F. E. L., Wolski, P., Lehner, F., Tebaldi, C., van Oldenborgh, G. J., Hogesteeger, S., Singh, R., Holden, P., Fučkar, N. S., Odoulami, R. C., & New, M. (2018). Anthropogenic influence on the drivers of the Western Cape drought 2015–2017. *Environmental Research Letters*, *13*(12), 2015–2017. https://doi.org/10.1088/1748-9326/aae9f9

Owen, N. A., Fahy, K. F., & Griffiths, H. (2015). Crassulacean acid metabolism (CAM) offers sustainable bioenergy production and resilience to climate change. *GCB Bioenergy*, *8*(4), 737–749. https://doi.org/10.1111/gcbb.12272

Palgrave, C. K. (1977). *Trees of southern Africa*. C Struik Publishers.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*, 181–197. https://doi.org/10.1890/07-2153.1

Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*, 1126–1136. https://doi.org/10.1111/2041-210X.12397

Qin, Z., Zhang, J. E., Jiang, Y. P., Wang, R. L., & Wu, R. S. (2020). Predicting the potential distribution of Pseudomonas syringae pv. actinidiae in China using ensemble models. *Plant Pathology*, *69*, 120–131. https://doi.org/10.1111/ppa.13112

Raes, N., & Aguirre-Gutiérrez, J. (2018). *A modeling framework to estimate and project species distributions in space and time* (pp. 309–320). Mountains.

Ringelberg, J. J., Zimmermann, N. E., Weeks, A., Lavin, M., & Hughes, C. E. (2020). Biomes as evolutionary arenas: Convergence and conservatism in the trans-continental succulent biome. *Global Ecology and Biogeography*, *29*, 1100–1113. https://doi.org/10.1111/geb.13089

Rödder, D., Schmidtlein, S., Veith, M., & Lötters, S. (2009). Alien invasive slider turtle in unpredicted habitat: a matter of niche shift or of predictors studied? *PLoS One*, *4*, e7843. https://doi.org/10.1371/journal.pone.0007843

RStudio Team (2019). RStudio: Integrated Development for R Studio Inc. http://www.rstudio.com

Saupe, E. E., Barve, V., Myers, C. E., Soberón, J., Barve, N., Hensz, C. M., Peterson, A. T., Owens, H. L., & Lira-Noriega, A. (2012). Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling*, *237–238*, 11–22. https://doi.org/10.1016/j.ecolmodel.2012.04.001

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS One*, *8*(8), e71218. https://doi.org/10.1371/journal.pone.0071218

Smith, S. D., Monson, R., & Anderson, J. E. (2012). *Physiological ecology of North American desert plants*. Springer.

Soberón, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 19644–19650. https://doi.org/10.1073/pnas.0901637106

Stock, W. D., Allsopp, N., van der Heyden, F., & Witkowski, E. T. F. (1997). In R. M. Cowling, D. M. Richardson, & S. M. Pierce (Eds.), *Plant form and function. In Vegetation of Southern Africa* (p. 615). Cambridge University Press.

Thuiller, W., Georges, D., & Engler, R. (2014). Biomod2: ensemble platform for species distribution modeling. *R Package Version*, *3*, 1–64.

Thuiller, W., Lavorel, S., & Araújo, M. B. (2005). Niche properties and geographical extent as predictors of species sensitivity to climate change. *Global Ecology and Biogeography*, *14*, 347–357. https://doi.org/10.1111/j.1466-822X.2005.00162.x

Title, P. O., & Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, *41*, 291–307. https://doi.org/10.1111/ecog.02880

Trabucco, A., & Zomer, R. J. (2018). Global Aridity Index and Potential Evapo-Transpiration (ET0) Climate Database v2 Methodology and Dataset Description.

Václavík, T., & Meentemeyer, R. K. (2009). Invasive species distribution modelling (iSDM): are absence data and dispersion constraints needed to predict actual distributions? *Ecological Modelling*, *220*, 3248–3258.

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*, 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

von Willert, D. J., Eller, B. M., Werger, M. J. A., Brinckmann, E., & Ihlenfeldt, H.-D. (1992). *Life Strategies of Succulents in Deserts: With Special Reference to the Namib Desert*. Cambridge University Press.

Webb, D. B., Wood, P. J., Smith, J. P., & Henman, G. S. (1984). *A guide to species selection for tropical and sub-tropical plantations*. Commonwealth Forestry Institute, University of Oxford.

Wilson, A. M., & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLOS Biology*, *14*(3), e1002415. https://doi.org/10.1371/journal.pbio.1002415

Winter, K. (1985). Crassulacean acid metabolism. In J. Barber, & N. R. Baker (Eds.), *Photosynthetic Mechanisms and the Environment* (pp. 329–387). Elsevier.

Winter, K., & Smith, J. A. C. (1996). *Crassulacean acid metabolism: Biochemistry, ecophysiology and evolution*. Springer-Verlag.

Wisz, M. S., & Guisan, A. (2009). Do pseudo-absence selection strategis influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, *9*, 1–3.

Yan, X., Tan, D. K. Y., Inderwildi, O. R., Smith, J. A. C., & King, D. A. (2011). Life cycle energy and greenhouse gas analysis for agave-derived bioethanol. *Energy and Environmental Science*, *4*, 3110–3121. https://doi.org/10.1039/c1ee01107c

Yang, W., Ma, K., & Kreft, H. (2013). Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *Journal of Biogeography*, *40*, 1415–1426. https://doi.org/10.1111/jbi.12108

Yao, A. Y. M. (1974). Agricultural potential estimated from the ratio of actual to potential evapotranspiration. *Agricultural Meteorology*, *13*, 405–417. https://doi.org/10.1016/0002-1571(74)90081-8

Zizka, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*, 744–751. https://doi.org/10.1111/2041-210X.13152

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Buckland, C. E., Smith, A. J. A. C., & Thomas, D. S. G. (2022). A comparison in species distribution model performance of succulents using key species and subsets of environmental predictors. *Ecology and Evolution*, *12*, e8981. https://doi.org/10.1002/ece3.8981