Robustness and accuracy of Maxent niche modelling for *Lactuca* species distributions in light of collecting expeditions

short title: Maxent model and Lactuca collecting expedition

MMP Cobben*[1,4], R van Treuren[1], NP Castañeda-Álvarez[2], CK Khoury[2,3], C Kik[1], TJL van Hintum[1]

*corresponding author, e-mail: m.cobben@nioo.knaw.nl

[1] Centre for Genetic Resources, The Netherlands (CGN)

P.O. Box 16

6700 AA Wageningen

The Netherlands

[2] International Center for Tropical Agriculture (CIAT)

Apartado Aéreo 6713

Calí, Colombia

[3] Centre for Crop Systems Analysis, Wageningen University

P.O. Box 16

6700 AA Wageningen

The Netherlands

[4] Netherlands Institute of Ecology (NIOO-KNAW)

P.O Box 50

6700 AB Wageningen

The Netherlands

ABSTRACT

Niche modelling software can be used to assess the probability of detecting a population of a plant species at a certain location. Here we used the distribution of wild relatives of lettuce (*Lactuca spp.*) to investigate the applicability of Maxent species distribution models for collecting missions. Geographic origin data of genebank and herbarium specimens and climatic data of the origin locations were used as input. For *L. saligna* we varied the input data by omitting specimens from different parts of the known distribution area to assess the robustness of the predicted distributions. Furthermore, we examined the accuracy of the modelling by comparing the predicted probabilities of population presence against recent expedition data for the endemic *L. georgica* and the cosmopolitan *L. serriola*. We found Maxent to be quite robust in its predictions, although its usefulness was higher for endemic taxa compared to more widespread species. The exclusion of occurrence data from the perceived range margins of the species can result in important information about local adaptation to distinct climatic conditions. We discuss the potential for enhanced use of Maxent in germplasm collecting planning.

Key words: Genebank, collecting expedition, *Lactuca*, species distribution modelling, niche modelling, Maxent, plant genetic resources

INTRODUCTION

Species distribution models have been used to predict distributions of a wide range of taxa (e.g. Guisan and Thuiller, 2005; Araújo and Guisan, 2006), and are increasingly employed for crop wild relatives (Guarino, 1995; Afonin and Greene, 1999; Greene *et al*., 1999; Jarvis *et al*., 2005; Maxted *et al*., 2008; Parra-Quijano *et al*., 2012). The Crop Wild Relatives & Climate Change project of the Global Crop Diversity Trust and Millennium Seed Bank Partnership, Royal Botanic Garden, Kew (www.cwrdiversity.org) uses a gap analysis methodology developed by Ramírez-Villegas *et al.* (2010), which incorporates the use of the maximum entropy model "Maxent" (Phillips *et al.*, 2006) to support collection planning for crop wild relatives. The input data for Maxent include the geographic origins of both genebank and herbarium specimens, and statistics representing the current climate, i.e. a set of temperature and precipitation parameters. The Maxent output distribution maps intend to give an indication of locations where the species may be present.

Collectors of plant genetic resources (PGR) are interested in material with new genetic diversity, preferably from species that can be crossed with the cultivated taxa for crop improvement. Such PGR can typically be collected from regions that have not been sampled previously. The gap-analysis method for crop genepools (Ramírez-Villegas *et al.*, 2010) assesses the priority with which a certain crop wild relative should be collected using the sampling representativeness score (SRS), estimating gross representation in genebanks; and the geographic representativeness score (GRS) and the environmental representativeness score (ERS), of which the GRS estimates comprehensiveness of genebank collections regarding the geographic range of the species, and ERS estimates comprehensiveness based on a principal component analysis (PCA) of the full environmental range of the modelled species. As such, it is possible to get an overview of populations that are underrepresented in genebanks and which may contain novel genetic diversity. The GRS and ERS depend upon

Maxent distribution models. In the method, it is implicitly assumed that the herbaria data provide full coverage of the distribution area of the species. In practice, herbaria data are incomplete and sampling bias can result in the systematic exclusion of certain regions of the species distribution from presence data. As a result, the distribution deduced with an associative species distribution model such as Maxent is not necessarily the complete distribution of the species. Maxent improves its models by excluding part of the presence data from the training sample to subsequently use it as test sample. However, this test sample is selected randomly, so does not systematically exclude a specific area, which mimicks the detection of an undiscovered region of the species distribution.

Here we use the distribution of wild relatives of lettuce (*Lactuca spp.*) as a case study to investigate how the predicted distribution of *L. saligna* depends on the input occurrence data. In addition, we compare the Maxent distribution predictions against expedition data of the endemic *L. georgica* and the cosmopolitan *L. serriola.* The results are utilized to discuss the applicability of Maxent to support PGR collecting missions.

MATERIAL AND METHODS

Study material

The wild relatives of lettuce were chosen as the focus of the case study due to high available knowledge and data for this crop, and the coincidence with a *Lactuca* collecting mission in the Trans Caucasus organised by the Centre for Genetic Resources, the Netherlands (CGN) in 2013. Moreover, lettuce relatives represent a wide variety of both niche and distribution sizes, ranging from pan-temperate distributions for *L. serriola* (D'Andrea *et al.*, 2009; Alexander, 2013) and *L. saligna,* to the narrow endemism of *L. georgica* in the Caucasus region(Zohary, 1991) and thus permit an evaluation of modelling methods for a variety of species types.

We collected information about known *Lactuca* populations from both herbarium and genebank databases (Table 1). The International Lactuca Database (ILDB), Eurisco and the National Plant Germplasm System (NPGS) were used as the sources of information about genebank accessions. For herbarium samples we consulted the Global Biodiversity Information Facility (GBIF) as well as obtained occurrence records directly from herbaria and researchers (Supplement Table S1). We used the inventory by Van Treuren *et al.* (2012) to select the species known to belong to the genepool of cultivated lettuce (*L. sativa)* (Table 1) and to check for synonyms (Supplement Table S2) and reassignments from other genera. Van Treuren et al. (2012) did a survey of the International Plant Names Index, which revealed a total number of 538 Lactuca species, of which 357 referred to synonyms and basionyms, whereas for another 51 the taxonomic status and their belonging to the genus Lactuca was questionable. Of the remaining 130 species, 20 are generally considered to be part of the lettuce gene pool (Table 1).

The data were cleaned, records without and with only coarse geographic information were removed. In addition, we removed the duplicate species-specific locations. For the

remaining locations we used DIVA-GIS (vs 7.5, www.diva-gis.org) (Hijmans et al 2012) to cross-check the match between longitude/latitude combination and the stated country.

Species distribution modelling

Current climatic conditions were downloaded from www.worldclim.org at a scale of 2.5 arc minutes, including 19 bioclimatic variables (Table 2). Maxent (Phillips *et al.*, 2006) uses a presence-only dataset as inputs and background points as pseudo-absences. The required 10,000 background points were selected randomly in each of the continents with species-specific occurrence data to ameliorate sample bias (VanDerWal *et al.*, 2009). Populations which were located outside the grid of climate cells, e.g. in the sea, were removed from the dataset.

Robustness of predicted distributions

L. saligna was chosen to assess the robustness of the Maxent projections in the marginal areas of the distribution range, since this species has a pan-temperate distribution with sufficient data points to exclude specific regions. Together with *L. serriola* and *L. virosa*, *L. saligna* serves as an important source of novel diversity for exploitation in the development of novel lettuce varieties by breeding companies (Lebeda *et al*., 2009).

Occurrence locations of *L. saligna* included North America, Australia, and Europe and from there extending into the eastern Mediterranean and Caucasus (Supplement Figure S1). To assess the robustness of the projections, we excluded the *L. saligna* occurrence samples in, respectively, the Greek region (GRC, 88 occurrences), the Israeli region (ISR, 32 occurrences) and the Eurasian region (EUR, 684 occurrences) (Supplement Figure S1). The Eurasian region was excluded to serve as a benchmark for the other predictions. The Greek and Israeli regions are very distinct marginal areas within the Eurasian distribution, from a

geographic point of view (Supplement Figure S1). Maxent was run for each of the three new datasets and the output was compared with the model based on the original, complete dataset. We used a ten fold division of the input data, each fold replicating the model using the consecutive parts as the test sample, while the remaining 90% of the input data was used as the training sample. For visual comparison we used the Maxent projections, based on this tenfold cross-validation for each dataset. In addition, we correlated the Maxent estimated probabilities of occurrence of each 2.5 arc minute cell in each of the excluded regions that resulted from the models with the reduced and the original dataset. This indicates how the estimated probability of occurrence in a single cell changes when the occurrence data is excluded from the dataset on which the probability model was based.

Relationship with expedition data

A *Lactuca* expedition to the Trans Caucasus was organised by CGN in 2013, which resulted in 94 unique collection locations. We compared the presence and absence locations of *L. georgica* and *L. serriola* with the model predictions for these locations. The model predictions were made *without including* the data of the sites visited during the expedition. So, while the modelling was done after the completion of the expedition, we mimicked the data availability prior to the collecting expedition, as if the modelling was done in its preparation. *L. georgica* and *L. serriola* were chosen because these 2 species were collected in a fair number of populations (32 and 55, respectively). In addition, they represent two opposites of the endemism spectrum, with *L. georgica* being endemic to the Trans Caucasus, while *L. serriola* has a pan-temperate distribution, with very many data points (Table 1). To investigate the effect of zooming in on the target area, a new *L. serriola* model was made using only the 133 known occurrence samples from the expedition region, again without including data from the newly sampled populations. The predicted probability of occurrence was determined for the

160 2.5 arc minute grid cell in which a population was found. All probabilities were classified in
161 categories ranging from 0.1 to 1.0, and the number of presence and absence locations were
162 summed per class.
163

RESULTS

Quality of the models

The four Maxent models for *L. saligna* (the full dataset and those excluding respectively GRC, ISR and EUR) and the model for *L. georgica* classified as valid models (Supplement Table S3) according to the gap-analysis protocol (Ramírez-Villegas *et al.*, 2010), for which the average test AUC (area under the curve) should be larger than 0.7, the standard deviation of the test AUC smaller than 0.15 and the proportion of the potential distribution area with a standard deviation of the estimated probability of occurrence > 0.15 should be smaller than 10%. The *L. serriola* model including all occurrence data was not considered valid, as a result of the low average test AUC of 0.65. However, when the analysis was limited to the expedition region this resulted in a valid model for *L. serriola*.

Robustness of model predictions for *L. saligna*

For *L. saligna*, excluding the occurrence data points in the Greek region of investigation (GRC) resulted in small changes in the estimated probabilities of occurrence in this region. The patterns in probability distributions were very similar between the two models (Fig. 1 maps) with a good correlation between both models (Fig. 1 scatter plot) with a small decrease in estimated probabilities in the model where the occurrences in GRC were omitted as compared to the model that included all occurrences.

Excluding the occurrence data points in the Israeli region of investigation (ISR) led to large changes in the estimated probabilities of occurrence of *L. saligna* in this region (Fig. 2). The similarity in the patterns of probability distributions was recognisable (Fig. 2 maps), but the decrease in estimated probabilities from the model in which the local occurrence data were omitted was substantial (Fig. 2 scatter plot).

When all the occurrences in the Eurasian region of investigation (EUR) were excluded, the probabilities of occurrence of *L. saligna* changed drastically (Fig. 3 scatter plot), showing increases and decreases depending on the location. The pattern of the potential distribution changed from an emphasis on Western Europe to a most probable occurrence in the Middle-East and Central Asia (Fig. 3 maps). The maximum estimated probability of occurrence increased from 0.75 to 0.97. However, the margins of the potential distribution of *L. saligna* were very similar between both models.

Model predictions in relation to expedition data

Fig. 4 shows the presence and absence of both *L. georgica* and *L. serriola* in 94 unique locations sampled during the CGN *Lactuca* expedition in 2013 in relation to the Maxent estimated probabilities of occurrence in the matching 2.5 arc minute grid cells. The model performed quite well for *L. georgica*, as at locations with low estimated probabilities of occurrence only absence of the species was observed, while at the locations with high estimated probabilities of occurrence observed presence was considerably higher than absence of the species . In contrast, the *L. serriola* projection showed very little differentiation across the expedition region, restricting the estimated probability of occurrence to only a few classes in the middle of the potential range. The majority of locations fell within the 0.4-0.5 probability class, at which more or less equal numbers of absences and presences were observed. The new *L. serriola* model, excluding all occurrence data but the ones in the expedition region, showed much more differentiation. However, the estimated probability of occurrence appeared a poor predictor for the presence and absence of the species.

213 **DISCUSSION**

214

215 **Species distribution modelling**

216 Species distribution models have been used for a number of decades and for many purposes
217 (review by Elith and Leathwick, 2009), such as the consequences of climate change on
218 species distributions and the assessment of the distribution of an invasive species. With the
219 growing availability of digital records from natural history museums, herbaria, and genebanks
220 coupled with the demand for mapped predictions, the incentive to put this source of presence-
221 only data to use has been increasing (Elith and Leathwick, 2009). Many different modelling
222 techniques exist and the discussion on how to best model presence-only data continues. It is
223 now common to compare presence data with background or pseudo-absence data, by e.g.
224 using regression methods such as generalised linear models (GLM), generalised additive
225 models (GAM) or multivariate adaptive regression splines (MARS), but also GARP (Genetic
226 Algorithm for Rule Set Production, Stockwell and Peters, 1999), ENFA (Ecological Niche
227 Factor Analysis, Hirzel *et al.*, 2002) and Maxent (Phillips *et al.*, 2006), in chronological order.
228 Elith *et al.* (2006) review and compare these methods. More recently, a platform was
229 developed to combine different techniques for ensemble forecasting of species distributions
230 (Araújo and New, 2007), called BIOMOD (R-package, http://R-Forge.R-project.org) (Thuiller
231 *et al*., 2009). This software is able to fit and compare different model classes in an attempt to
232 reach more robust forecasts by treating the methodological uncertainties in different models.
233 The latest version of BIOMOD, biomod2, also includes Maxent as one of the techniques.

234 In this study we have chosen to use Maxent as the modelling technique to align our
235 results with the gap analysis method (Ramírez-Villegas *et al.*, 2010), which is used as input
236 for many planned collecting expeditions in the Crop Wild Relatives & Climate Change
237 project. With this project in mind, we have aimed to provide information about Maxent's

robustness and accuracy with regard to such collecting expeditions. Apart from these considerations, the accessibility and relative ease of use of the Maxent software compared to the others mentioned are valuable assets for application by non-experts in the planning of collecting expeditions. However, for increasing the strength of the predictions, using and comparing different modelling techniques as is done in BIOMOD is likely a valuable contribution. Compared to Maxent, this does require increased need for processing power, technical knowledge and time, which might limit the usability of BIOMOD for the non-expert.

Robustness of *L. saligna* models

From the comparison of the different *L. saligna* models, we conclude that the Greek region is not climatically distinct within the known *L. saligna* distribution area (Fig. 1). In contrast, the Israeli region appeared very distinct (Fig. 2). Comparing the response curves of the Maxent predictions to the different environmental variables (Fig. S2), we found the Israeli region to differ from other areas within the distribution range especially regarding the mean annual temperature (BIO1), the maximum temperature of the warmest month (BIO5), the variation in the precipitation over the seasons (BIO15), and the total precipitation of the warmest quarter (BIO18). Fig. 2 indicates that the model is not capable of predicting this latter region as a potential distribution area when data from this region are excluded, while the unique climatic conditions may indicate the presence of potentially interesting diversity. For the Eurasian region of investigation, although the estimated local probabilities differ substantially between the two models (Fig. 3), the distribution borders are very similar. The changes in distribution pattern indicate that the south-eastern region, where the probabilities in occurrence increase when the Eurasian occurrences are excluded, is climatically more similar to the other regions in the world where *L. saligna* is sampled. From this it follows that, in addition to the Israeli

region, Western Europe may be considered a relatively exceptional climatic region in the *L. saligna* distribution area.

From this analysis we conclude that Maxent's outputs are generally robust, yet on a local scale, on which a collecting mission is typically planned, estimated probabilities of occurrence can differ to a larger extent. This depends on whether populations in the specific region have been sampled before (and the data used as input for the model), and on the climatic relatedness to other sampled regions within the distribution. It is important to note that all species distribution models project suitable habitat by climate association, and thus that none can predict potential distribution in regions that are climatically unique compared to regions where the species has been sampled. Particularly for the purpose investigated here, in search of unique plant genetic resources, this is an important limitation. However, excluding occurrence data from the perceived range margins of the species may result in important information about local adaptation to distinct climatic conditions. A principal component analysis of the climatic data from the occurrence locations can provide the same information and may be a good starting point for such an analysis. There are other possibilities to get information about local adaptation, e.g using SNP (single nucleotide polymorphism) data to confirm that a population is genetically different from similar populations, or checking phenotypic characteristics. Such methods tend to be used at a later stage, while the method we suggest here can be done with currently available information about population locations to get an indication about where such useful locally adapted populations may be present.

L. saligna is native to Eurasia and North-Africa (GRIN, 2014; Lebeda *et al.*, 2004a; Lebeda *et al.,* 2004b), and when omitting the Eurasian occurrence locations, we have essentially predicted the native distribution of the species from its non-native distribution. While this is not a logical procedure for collection planning, it gives indication of robustness of the predictions and the possibility to predict species' presence in a region where the species

has not been sampled. Omitting many data points from the native area of the species then makes an interesting benchmark study with which to compare the other results from the Greek and Israeli regions. Interestingly, this procedure could be used to indicate the potential origin of populations in the non-native distribution area (Alexander, 2013).

Relationship with expedition data

L. georgica is an endemic (Zohary, 1991) species, that lives in an equilibrium environment where natural competition determines the distribution of species. *L. serriola* is a ruderal species, thus living in disturbed environments (Grime, 1977), with a cosmopolitan distribution (D'Andrea *et al.*, 2009; Alexander, 2013). The latter's widespread distribution and the very many data points that are included in the initial model (Table 1) likely account for the rather undifferentiated model projections observed for *L. serriola*. It is interesting to note that, although not very informative, this model does provide a good prediction of its occurrence, with a similar number of presence and absence locations across the 50% probability region. When we limited the model to the expedition region, the number of probability classes increased substantially. However, actual presence and absence locations appeared to correlate poorly with the corresponding probability of occurrence according to the modelling (Fig. 4c). The *L. serriola* absence locations are mostly *L. georgica* presence locations, representing fairly undisturbed habitats. In addition, the ruderal nature of *L. serriola*, combined with its global distribution, explains its relative insensitivity to climatic conditions. Thus, the Maxent model for this global, ruderal species would not have been informative or otherwise useful for the Trans Caucasus expedition, not even when the projections would have been restricted to the region of interest. This is in line with the gap analysis protocol as suggested by Ramírez-Villegas *et al*. (2010), who excluded weedy species from the analysis. Here it needs to be noted that the collection of wide-spread and ruderal species such as *L. serriola* does not

require any modelling support, since such species can be easily located. In contrast, expected and actual presence data correlated well for the endemic *L. georgica*, living in pristine habitat. In fact, a population was sampled in a region where the local experts had not expected it, while the model predicted a high probability of occurrence at this location. In the case that the projection for *L. georgica* would have been available prior to the collecting mission, the expedition route would have been slightly adjusted to explore a nearby region where also high probabilities of occurrence were estimated. Thus, Maxent distribution models may be useful to support collecting missions and based on our findings this may particularly apply to endemic species growing in relatively undisturbed habitats.

Concluding remarks

Based on our results, we recommend organisers of collecting missions to run Maxent or similar species distribution models for their species of interest prior to the expedition in complement to expert knowledge on species distributions. Given sufficient input data, particular faith may be given to the model results for endemic species amongst the range of relevant crop wild relatives. The resulting maps should be combined with the knowledge of local authorities to identify potential new populations of these species. In addition, excluding occurrence data from the perceived range margins of the species may result in important information about local adaptation to distinct climatic conditions.

To increase access to the methodology, avoiding the necessity of installing and operating the software, a web-based version of Maxent, including the worldclim.org climatic dataset and standard model settings, would greatly facilitate the application of species distribution modelling in the preparation phase of collecting missions, and would be particularly useful for plant genetic resource conservation efforts with limited resources.

ACKNOWLEDGMENTS

This study was part of the Fundamental Research Programme on Sustainable Agriculture (KB-12-005.03-003) funded by the Dutch Ministry of Economic Affairs. M.C. was additionally funded by the Open Programme of the Netherlands Organisation of Scientific Research (NWO). Data and methods were contributed by the Adapting Agriculture to Climate Change Project, which is managed by the Global Crop Diversity Trust with the Millennium Seed Bank Partnership of the Royal Botanic Gardens, Kew and supported by the Government of Norway. For further information see the project website: http://www.cwrdiversity.org/

We acknowledge the contribution of the breeding companies Agrisemen BV, Enza Zaden Research and Development BV, Monsanto Holland BV, Nunhems Netherlands BV and Rijk Zwaan Breeding BV to the *Lactuca* expedition in 2013.

REFERENCES

Afonin A and Greene SL (1999) Germplasm collecting using modern geographic information technologies: directions explored by the N.I. Vavilov Institute of Plant Industry. In: Greene SL and Guarino L (eds) *Linking Genetic Resources and Geography: Emerging Strategies for Conserving and Using Crop Biodiversity*. CSSA Special Publication 27. American Society of Agronomy, Madison, Wisconsin. pp.75–85.

Alexander JM (2013) Evolution under changing climates: climatic niche stasis despite rapid evolution in a non-native plant. *Proceedings of the Royal Society B* 280: 20131446.

Araújo MB and Guisan A (2006) Five (or so) challenges for species distribution models. *Journal of Biogeography* 33: 1677-1688.

Araújo MB and New M (2007) Ensemble forecasting of species distributions. *Trends in Ecology and Evolution* 22: 42-47.

D'Andrea L, Broennimann O, Kozlowski G, Guisan A, Morin X, Keller-Senften J and Felber F (2009) Climate change, anthropogenic disturbance and the northward range expansion of *Lactuca serriola* (Asteraceae). *Journal of Biogeography* 36: 1573-1587.

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129-151.

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

369 Greene SL, Hart TC and Afonin A (1999) Using geographic information to acquire wild crop
370 germplasm for *ex situ* collections: I. map development and field use. *Crop Science* 39: 836–
371 842.

372 Grime JP (1977) Evidence for the existence of three primary strategies in plants and its
373 relevance to ecological and evolutionary theory. *The American Naturalist* 111: 1169-1194.

374 GRIN (2014) http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?21359

375 Guarino L (1995) Mapping the ecogeographic distribution of biodiversity. In: Guarino L,
376 Ramanatha Rao V and Reid R (eds) *Collecting Plant Genetic Diversity Technical Guidelines.*
377 CAB International, Wallingford, UK. pp.287–327.

378 Guisan A and Thuiller W (2005) Predicting species distributions: offering more than simple
379 habitat models. *Ecology Letters* 8: 993-1009.

380 Hijmans RJ, Guarino L and Mathur P (2012) DIVA-GIS vs 7.5. A geographic information
381 system for the analysis of species distribution data. Software and manual available at
382 www.diva-gis.org.

383 Hirzel AH, Hausser J, Chessel D and Perrin N (2002) Ecological-niche factor analysis: How
384 to compute habitat-suitability map without absence data. *Ecology* 83: 2027-2036.

385 Jarvis A, Williams K, Williams D, Guarino L, Caballero PJ and Mottram G (2005) Use of
386 GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum Sendtn.*)
387 in Paraguay. *Genetic Resources and Crop Evolution* 52: 671–682.

388 Lebeda A, Dolezalová I, Feráková V and Astley D (2004a) Geographical distribution of wild
389 *Lactuca* species (Asteraceae, Lactuceae). *The Botanical Review* 70(3): 328-356.

390 Lebeda A, Dolezalová I, Astley D (2004b) Representation of wild *Lactuca* spp. (Asteraceae,
391 Lactuceae) in world genebank collections. *Genetic Resources and Crop Evolution* 51: 167-
392 174.

393 Lebeda A, Doležalová I, Křístková E, Kitner M, Petrželová I, Mieslerová B and Novotná A
394 (2009) Wild Lactuca germplasm for lettuce breeding: current status, gaps and challenges.
395 Euphytica 170:15–34.

396 Maxted N, Dulloo E, Ford-Lloyd BV, Iriondo JM, Jarvis A (2008) Gap analysis: a tool for
397 complementary genetic conservation assessment. *Diversity and Distributions* 14:1018–1030.

398 Parra-Quijano M, Iriondo JM, Torres E (2012) Improving representativeness of genebank
399 collections through species distribution models, gap analysis and ecogeographical maps.
400 *Biodiversity and Conservation* 21: 79-96.

401 Phillips SJ, Anderson RP and Schapire RE (2006) Maximum entropy modeling of species
402 geographic distributions. *Ecological Modelling* 190: 231-259.

403 Ramírez-Villegas J, Khoury C, Jarvis A, Debouck DG and Guarino L (2010) A gap analysis
404 methodology for collecting crop genepools: a case study with *Phaseolus* beans. *Plos ONE*
405 5(10): e13497.

406 Stockwell D and Peters D (1999) The GARP modelling system: problems and solutions to
407 automated spatial prediction. *International Journal of Geographical Information Science* 13:
408 143-158.

409 Thuiller W, Lafourcade B, Engler R and Araújo MB (2009) BIOMOD-a platform for
410 ensemble forecasting of species distributions. *Ecography* 32: 369-373.

411 Treuren R van, Coquin P and Lohwasser U (2012) Genetic resources collections of leafy
412 vegetables (lettuce, spinach, chicory, artichoke, asparagus, lamb's lettuce, rhubarb and rocket
413 salad): composition and gaps. *Genetic Resources and Crop Evolution* 59: 981-997.

414 VanDerWal J, Shoo LP, Graham C and Williams SE (2009) Selecting pseudo-absence data
415 for presence-only distribution modeling: how far should you stray from what you know?
416 *Ecological Modelling* 220: 589-594.

417 Zohary D (1991) The wild genetic resources of culitvated lettuce (*Lactuca sativa* L.).
418 *Euphytica* 53: 31-35.

419

420 **TABLES**

421

422 **Table 1.** The species of the primary (I), secondary (II) and tertiary (III) genepool of *Lactuca sativa* for

423 which we found occurrence data in herbaria and genebanks, respectively.

424

species	genepool	herbarium samples	genebank accessions
L. aculeata	I	3	4
L. altaica	I	3	2
L. azerbaijanica	I	0	0
L. dregeana	I	4	0
L. georgica	I	17	1
L. scarioloides	I	1	0
L. serriola	I	23520	1177
L. saligna	II	1451	102
L. virosa	II	3318	102
L. acanthifolia	III	34	0
L. aurea	III	0	0
L. longidentata	III	0	0
L. orientalis	III	141	0
L. quercina	III	106	6
L. sibirica	III	854	2
L. taraxacifolia	III	2	0
L. tatarica	III	861	12
L. viminea	III	728	12
L. watsoniana	III	0	0

425

426

427 **Table 2.** The 19 bioclimatic variables used as input for the model, downloaded at a scale of 2.5 arcmin
428 from www.worldclim.org
429

BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

430
431

FIGURE LEGENDS

Figure 1. The projected probabilities of occurrence for *L. saligna* in the Greek region when including and excluding the occurrence data of this region, on a scatter plot showing the changes in projected probabilities for the Greek region when omitting the occurrence data.

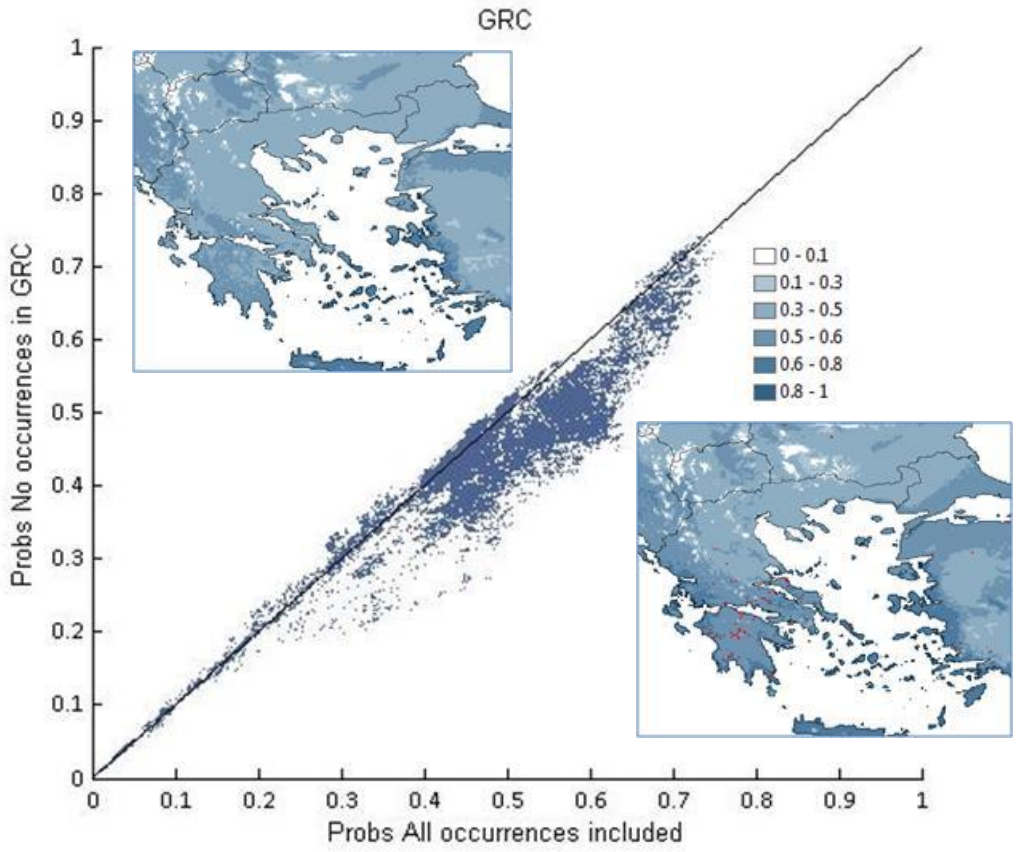
Figure 2. The projected probabilities of occurrence for *L. saligna* in the Israeli region when including and excluding the occurrence data in this region on a scatter plot showing the changes in projected probabilities for the Israeli region when omitting the occurrence data.

Figure 3. The projected probabilities of occurrence for *L. saligna* in the Eurasian region when including and excluding the occurrence data in this region on a scatter plot showing the changes in projected probabilities for the Eurasian region when omitting the occurrence data.

Figure 4. The number of locations in the region under study where *L. georgica* and *L. serriola* was present or absent plotted against the estimated probabilities of occurrence at these locations, based on the model including all occurrence data (*L. georgica* and *L. serriola*) and only occurrence data in the region under study (*L. serriola* CAU).

451 **FIGURES**

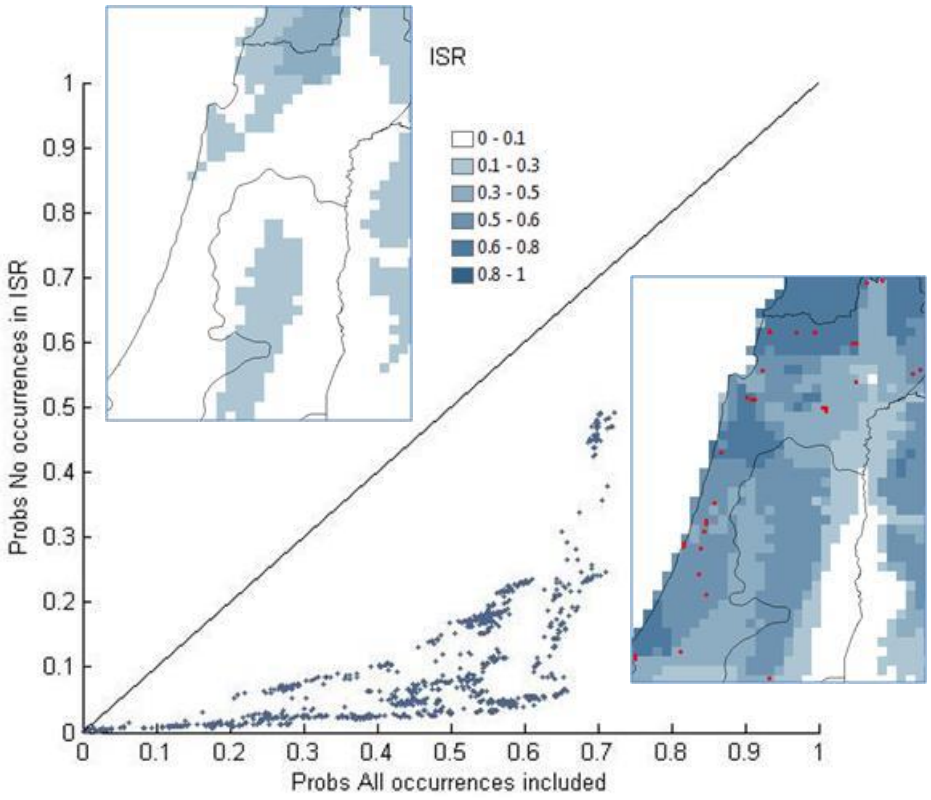
452 Figure 1



453

454

455 Figure 2

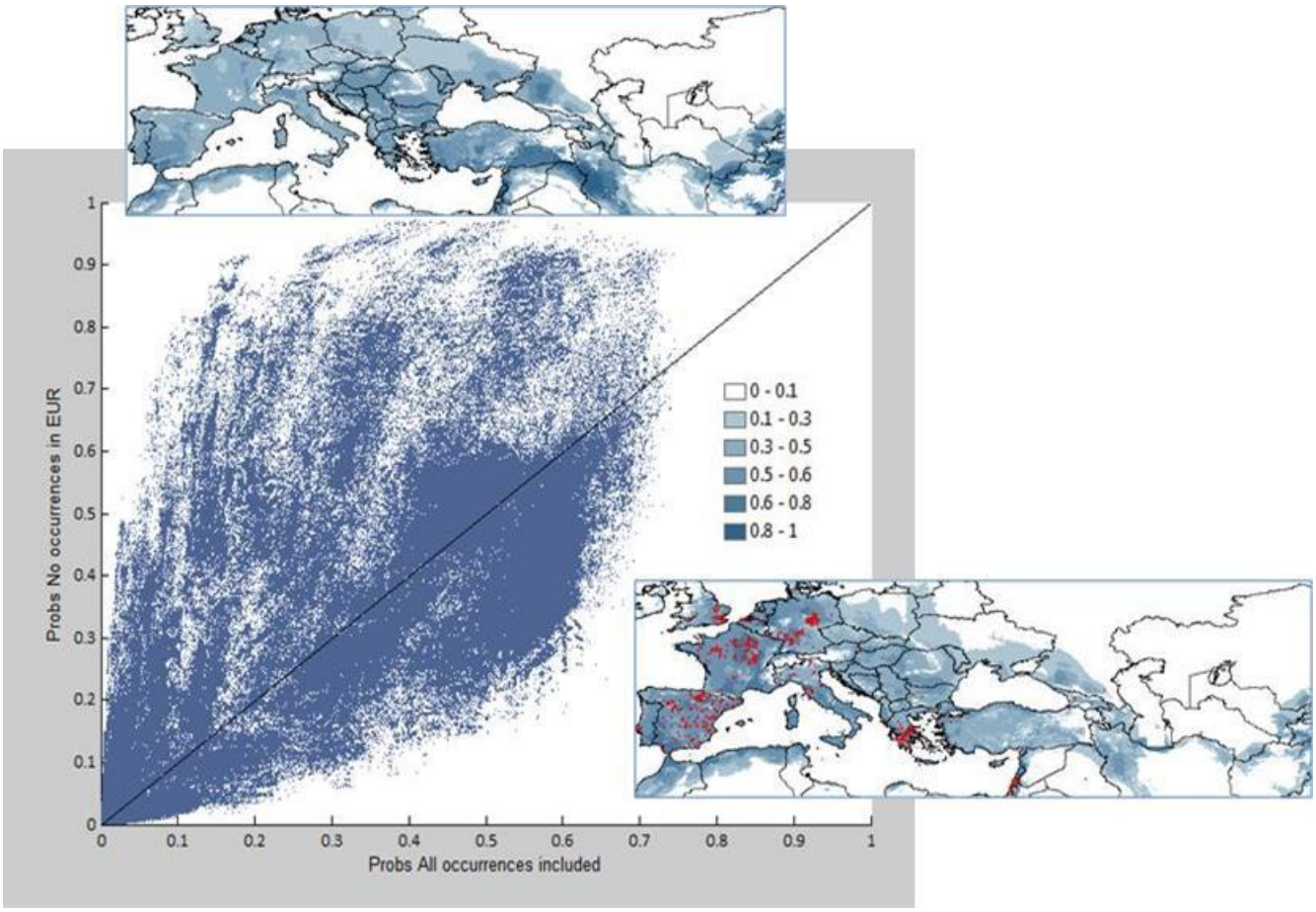


456

457

458

459 Figure 3



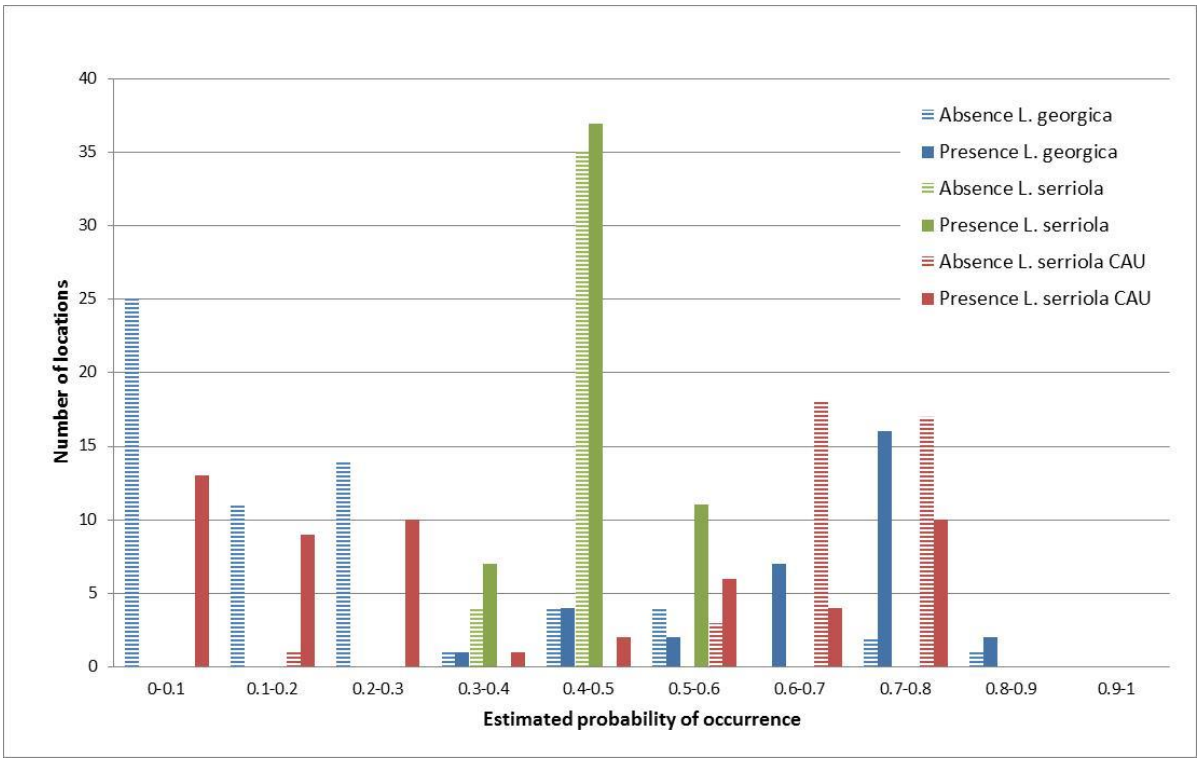
460

461

462

463

Figure 4


40
35
30
25
20
15
10
5
0
Number of locations
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
Estimated probability of occurrence
Absence L. georgica
Presence L. georgica
Absence L. serriola
Presence L. serriola
Absence L. serriola CAU
Presence L. serriola CAU

469 **SUPPLEMENT**

470 Table S1. Overview of all the herbaria and genebanks from which *Lactuca* samples were retrieved.

Academy of Natural Sciences Herbarium
Australia's Virtual Herbarium
Botanical Society of the British Isles Herbaria
California Academy of Sciences
Centro de Referência em Informação Ambiental
Consortium of Pacific Northwest Herbaria
Denver Botanic Gardens Herbarium
Eurisco
Field Museum
Florida State University Herbarium
Global Biodiversity Information Facility
Harvard University Herbarium
Instituto Superior de Agronomia
International Lactuca Database
Jardim Botanico do Rio de Janeiro
Manchester University Herbarium
Museum National d'Histoire Naturelle Herbarium
Nationaal Herbarium Nederland
New York Botanical Garden Herbarium
Real Jardin Botanico de Madrid
Royal Botanic Gardens Edinburgh
Royal Botanic Gardens, Kew
Smithsonian Institution, National Herbarium
United States Department of Agriculture, National Plant Germplasm System
Universidad del Valle Herbarium
Universidade Lisboa Museu Nacional de Historia Natural e da Ciência
University of California and Jepson Herbaria
University of California, Riverside Herbarium
University of Coimbra Herbarium
V. L. Komarov Botanical Institute
Wageningen University Herbarium
West Virginia University Herbarium
World Vegetable Center

471

Table S2. The species in the Lactuca genepool (Table 1) and their synonyms (Van Treuren *et al*., 2012). Author names were included only when they distinguish between different synonyms in the consulted databases (Table S1).

accepted names	synonyms
L. acanthifolia	*L. amorgina*
	L. eburnea
L. aculeata	
L. altaica	
L. aurea	
L. azerbaijanica	
L. dregeana	*L. virosa Thunb.*
L. georgica	
L. longidentata	*Scariola longidentata*
L. orientalis	
L. quercina	*L. altissima*
	L. armena
	L. chaixii
	L. sagittata
	L. stricta
	L. vialea
	L. wilhelmsiana
	L. cracoviensis
	L. cyanea
	L. decorticata
L. saligna	*L. adulteriana*
	L. angustifolia
	L. caucasica
	L. cracoviensis
	L. cyanea
	L. salicifolia
	L. spiciformis
	L. tommasiniana
	L. virgata
	L. virosa Habl.
	L. wallrothii
L. scarioloides	*L. kotschyana*
L. serriola	*L. albicaulis*
	L. augustana
	L. coriacea
	L. dubia
	L. latifolia
	L. plicata
	L. scariola
	L. sylvestris

	L. tephrocarpa
	L. virosa Luce
L. sibirica	
L. taraxacifolia	*L. alaica*
	L. kotschyi
	L. pentaphylla
L. tatarica	*L. clarkei*
	L. multipes
	L. oblongifolia
	L. pulchella (Pursh) DC.
	L. pulchella DC.
L. viminea	*L. alpestris*
	L. chondrilliflora
	L. decorticata
	L. numidica
	L. ramosissima
L. virosa	*L. agrestis*
	L. ambigua
	L. cornigera
	L. flavida
	L. lactucarii
	L. livida
	L. serratifolia
	L. sinuata
	L. virosa L.
L. watsoniana	

475

476

Table S3. The Maxent model statistics for each of the investigated species of the lettuce genepool. ATAUC: the 10-fold average test AUC (area under the curve), STAUC: the standard deviation of the test AUC of the 10 different folds, ASD15: the percentage of the potential distribution coverage with standard deviation above 0.15. For the gray colored species, the total number of samples is smaller than 10 (Table 1), meaning that the number of folds is equal to the number of samples.

Taxon	ATAUC	STAUC	ASD15	ValidModel
Lactuca acanthifolia	0.9984	0.0006	0	yes
Lactuca aculeata	0.9599	0.0667	0	yes
Lactuca altaica	0.8728	0.0571	44.195	no
Lactuca dregeana	0.9896	0.0061	4.433	yes
Lactuca georgica	0.9973	0.0043	0	yes
Lactuca orientalis	0.9736	0.0240	0.610	yes
Lactuca quercina	0.9810	0.0341	0.055	yes
Lactuca saligna	0.9183	0.0044	0	yes
Lactuca saligna_EUR	0.9422	0.0048	0.037	yes
Lactuca saligna_GRC	0.9208	0.0057	0	yes
Lactuca saligna_ISR	0.9191	0.0051	0	yes
Lactuca scarioloides	NA	NA	NA	no
Lactuca serriola	0.6490	0.0046	0	no
Lactuca serriola_TC	0.9895	0.0060	1.450	yes
Lactuca sibirica	0.9596	0.0038	0	yes
Lactuca taraxacifolia	0.5000	0	NA	no
Lactuca tatarica	0.9169	0.0061	0.126	yes
Lactuca viminea	0.9637	0.0043	0.003	yes
Lactuca virosa	0.8806	0.0042	0	yes

Figure S1a. The global region of analysis (ALL) and all *L. saligna* occurrences.

Figure S1b. The Eurasian region of analysis (EUR) and its *L. saligna* occurrences.

Figure S1c. The Greek region of analysis (GRC) and its *L. saligna* occurrences.

Figure S1d. The Israeli region of analysis (ISR) and its *L. saligna* occurrences.

Figure S1a

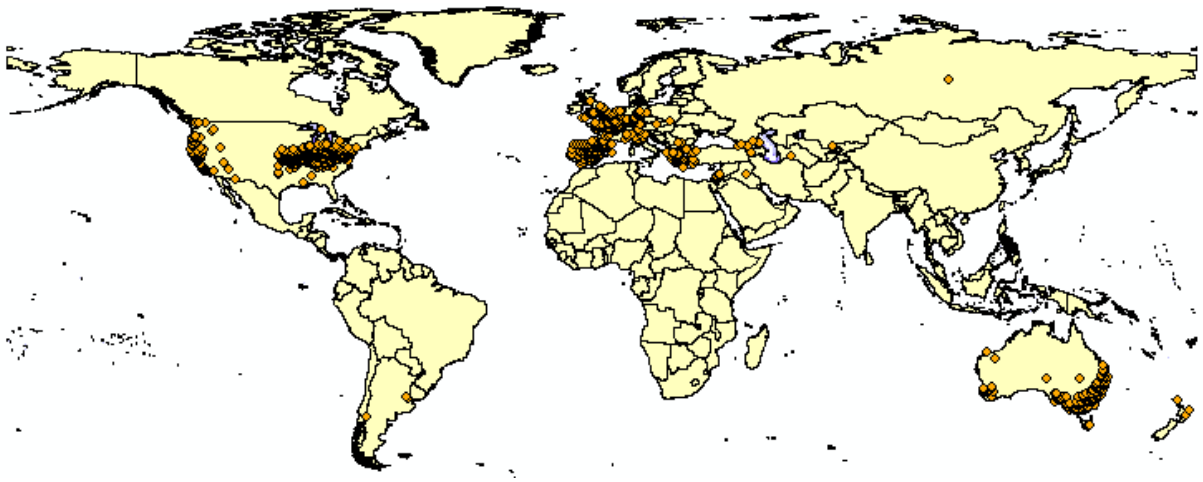


Figure S1b

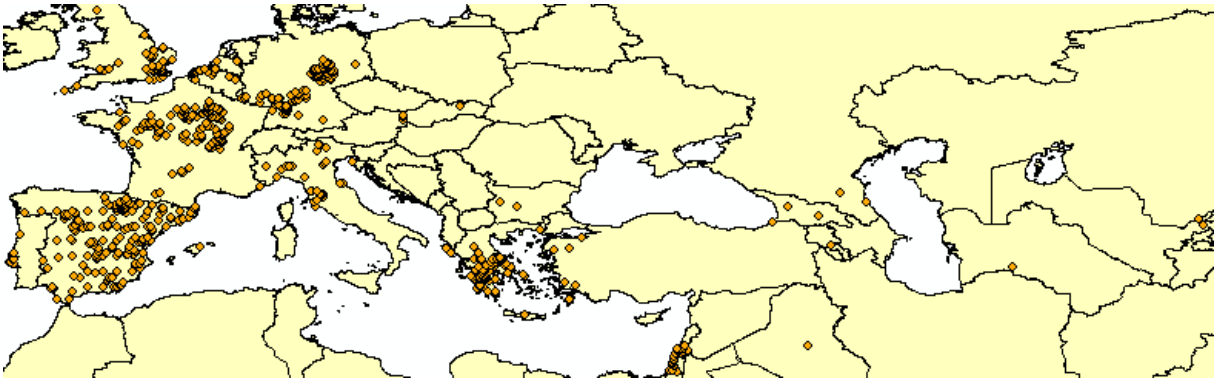


Figure S1c

Figure S1d

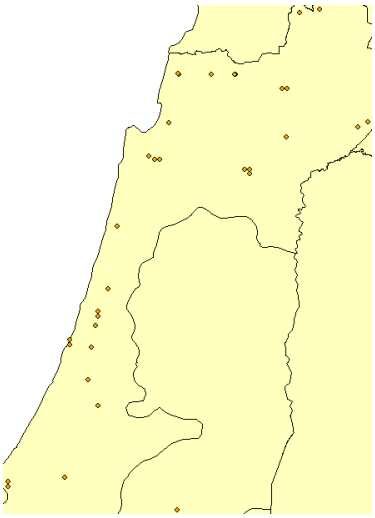


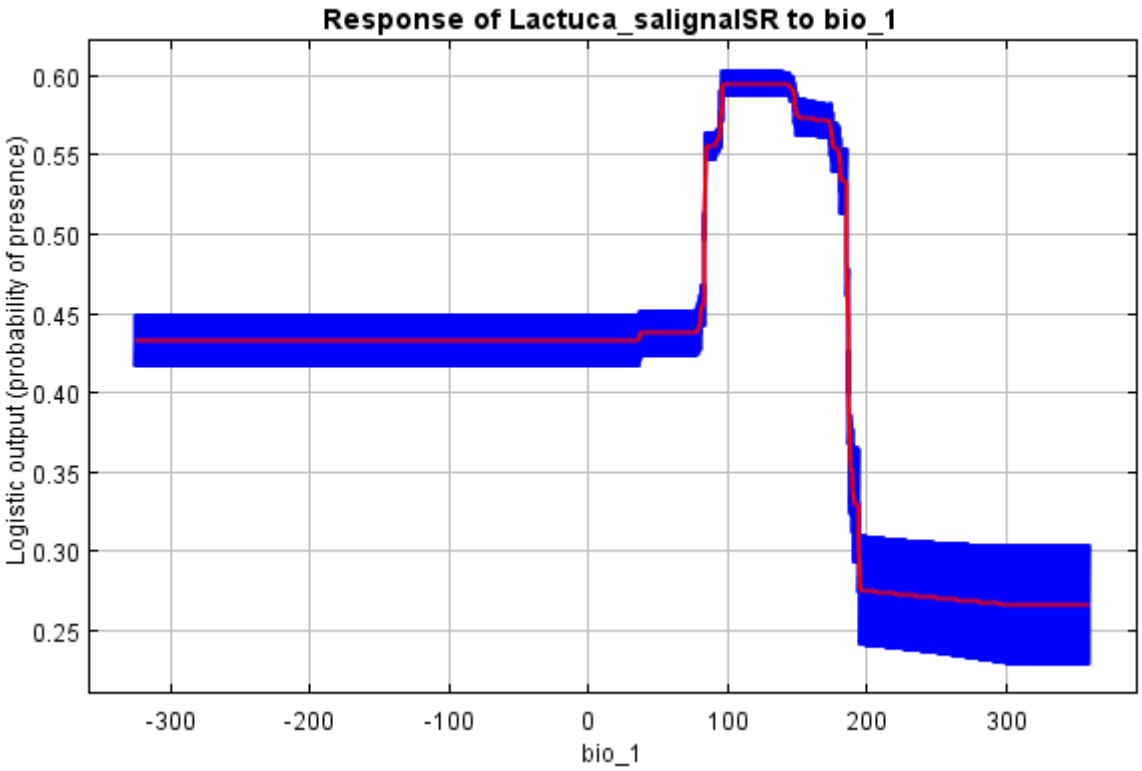
Figure S2a. The Maxent response curves to BIO_1 for the model based on all occurrences (model Lactuca_saligna) and the model for which the Israeli occurrences were omitted (model Lactuca_salignaISR). The curves show how the logistic prediction changes as the mean annual temperature is varied, keeping all other environmental variables at their average sample value.

Figure S2b. The Maxent response curves to BIO_5 for the model based on all occurrences (model Lactuca_saligna) and the model for which the Israeli occurrences were omitted (model Lactuca_salignaISR). The curves show how the logistic prediction changes as the maximum temperature of the warmest month is varied, keeping all other environmental variables at their average sample value.

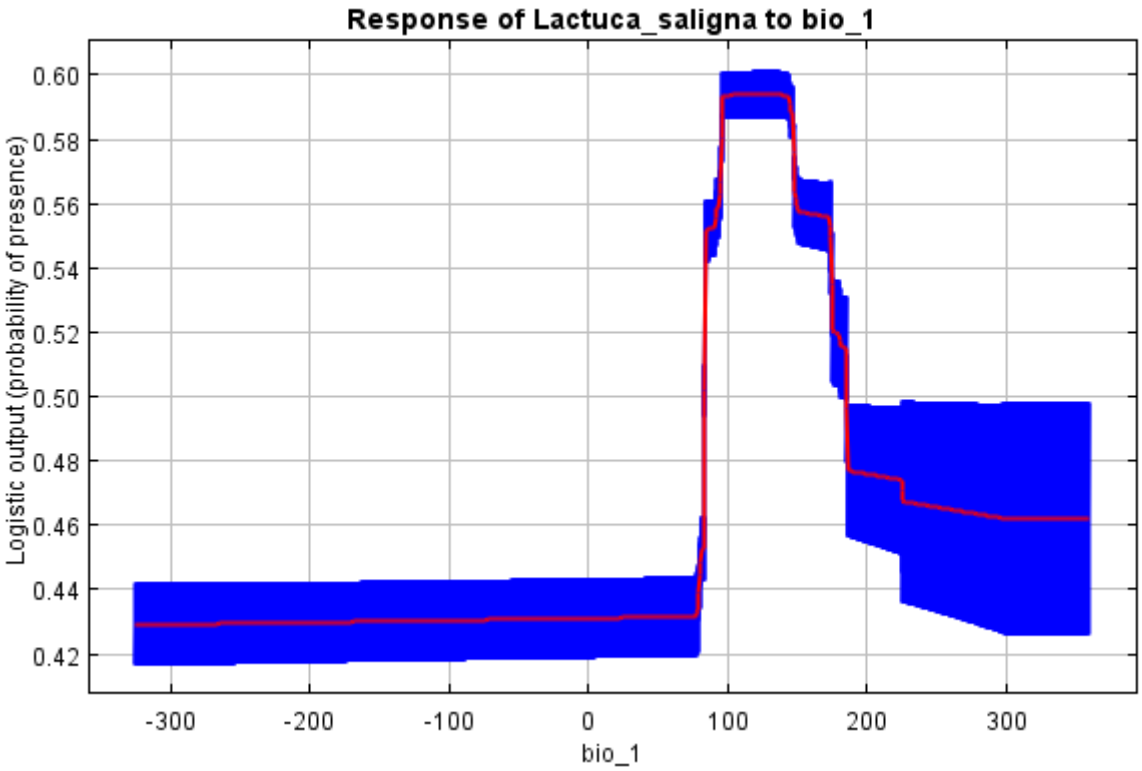
Figure S2c. The Maxent response curves to BIO_15 for the model based on all occurrences (model Lactuca_saligna) and the model for which the Israeli occurrences were omitted (model Lactuca_salignaISR). The curves show how the logistic prediction changes as the variation in the precipitation over the seasons is varied, keeping all other environmental variables at their average sample value.

Figure S2d. The Maxent response curves to BIO_18 for the model based on all occurrences (model Lactuca_saligna) and the model for which the Israeli occurrences were omitted (model Lactuca_salignaISR). The curves show how the logistic prediction changes as the total precipitation of the warmest quarter is varied, keeping all other environmental variables at their average sample value.

519 Figure S2a



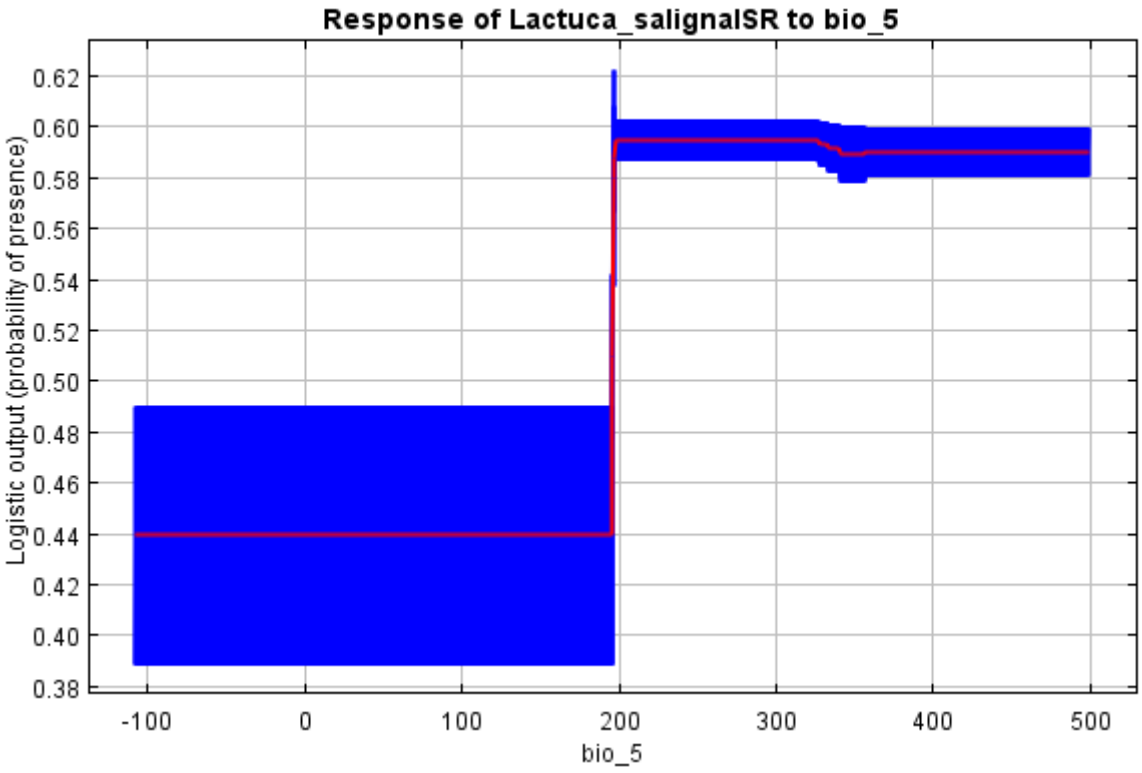
520



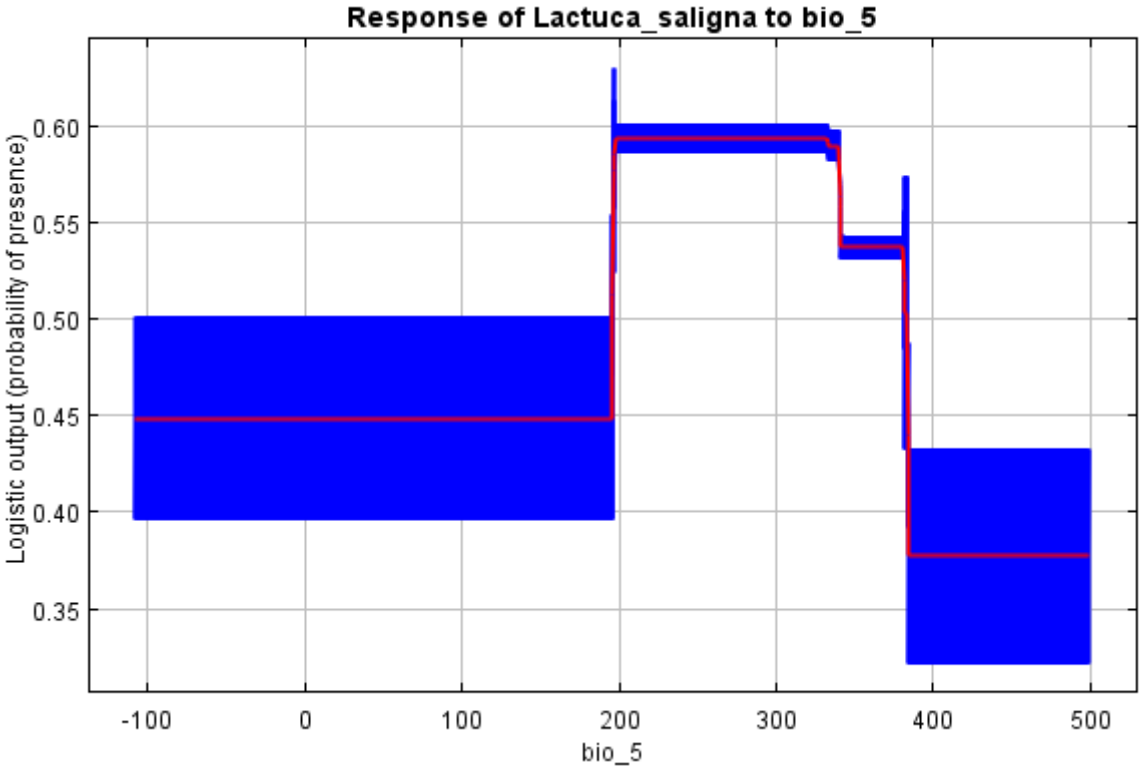
521

522

523 Figure S2b



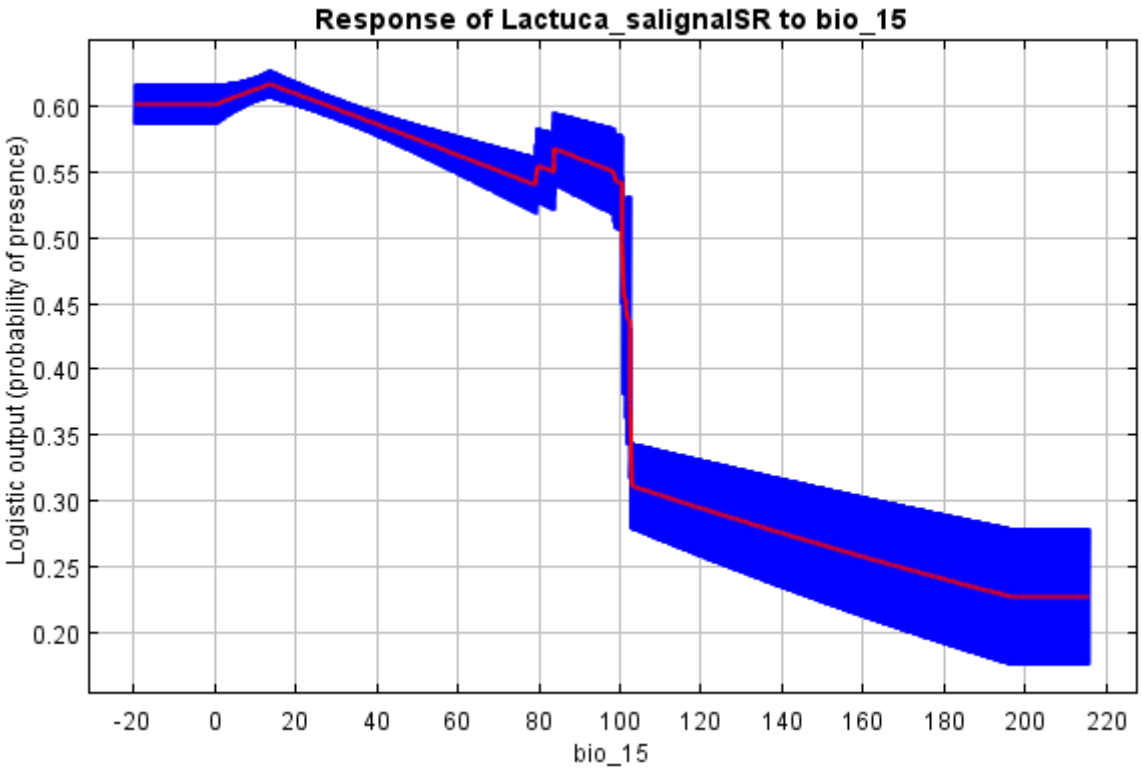
524



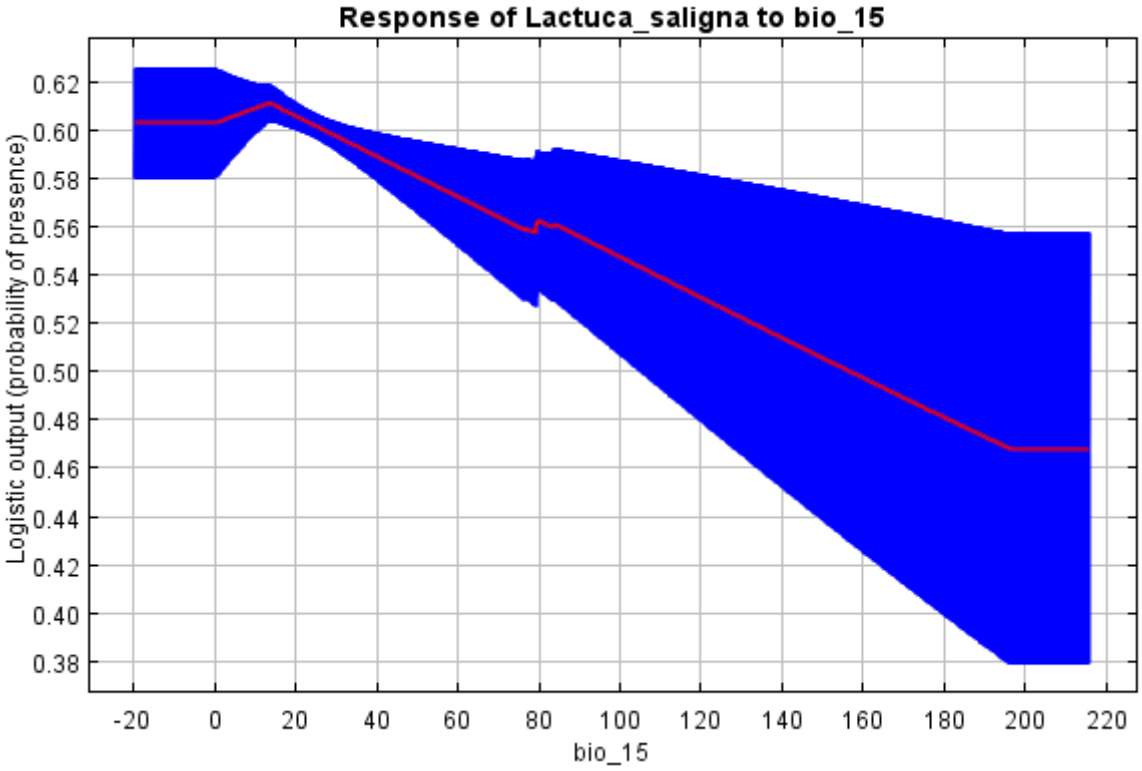
525

526

527 Figure S2c



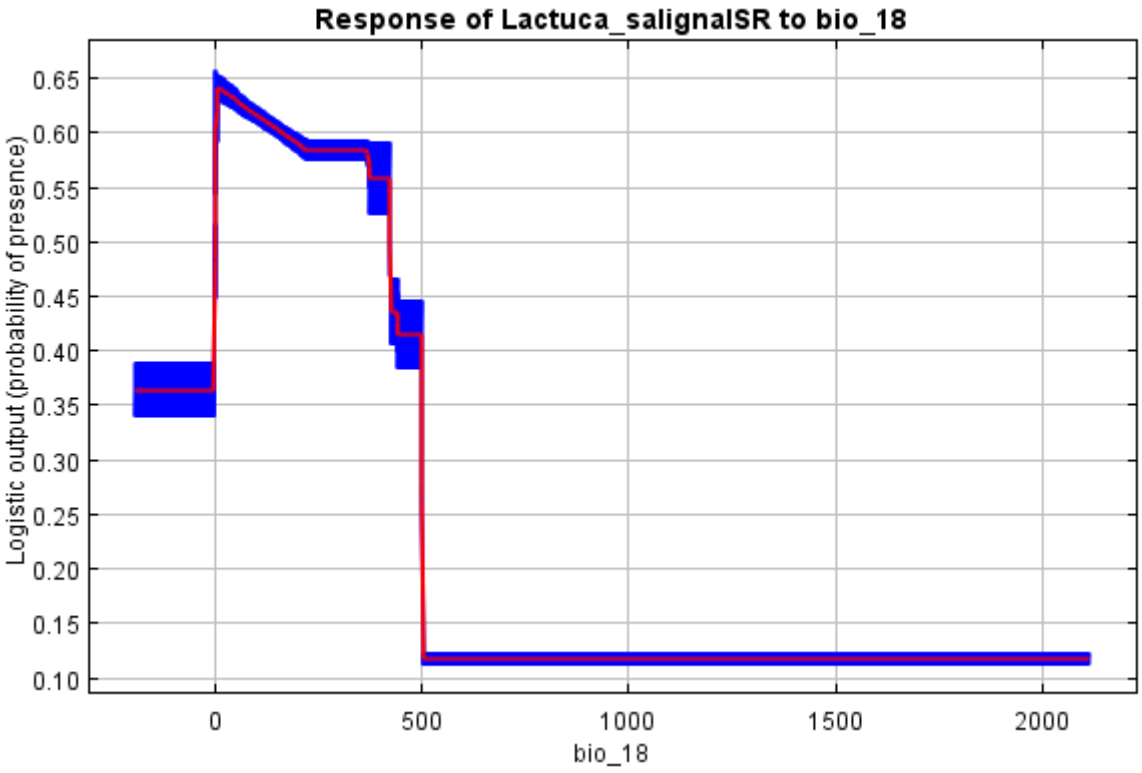
528



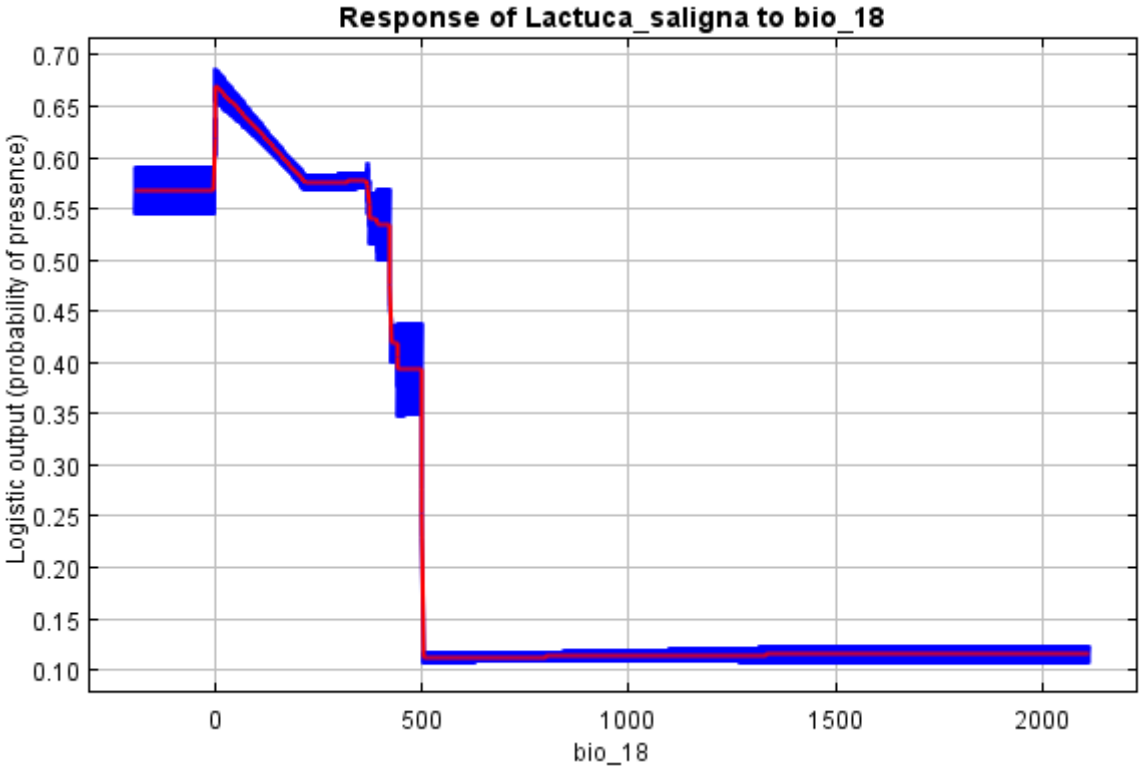
529

530

531 Figure S2d



532



533

534