THE INTERSECTION OF CLIMATE AND NICHE: LIKELIHOOD ESTIMATION OF MODERN AND PAST CLIMATE USING PLANT BIODIVERSITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Robert Simpson Harbert

August 2016

THE INTERSECTION OF CLIMATE AND NICHE: LIKELIHOOD ESTIMATION OF MODERN AND PAST CLIMATE USING PLANT BIODIVERSITY

Robert Simpson Harbert, Ph. D.

Cornell University 2016

Plant distributions have long been understood to be correlated with environmental conditions. Climate is one of the major components, and arguably the most important factor determining broad-scale species distributions. Therefore, it is expected that the complement of plants coexisting in a community is reflective of the local environment, particularly climate. A new method, Climate Reconstruction Analysis using Coexistence Likelihood Estimation (CRACLE), was developed during the course of this dissertation work to leverage the above hypothesis into a robust, quantitative approach for the estimation of climate from vegetation floristic data.

Based on test datasets and comparison with existing climate models generated from modern climate records, CRACLE produces highly accurate estimates of climate variables for which there are, or can easily be constructed, continuous model data for the Earth's land surfaces. It is shown that CRACLE can accurately estimate a

wide range of temperature, precipitation, evapotranspiration, and seasonal variables as generated from the WorldClim base variables. This represents an improvement on other comparable taxonomic methods (e.g., the Coexistence Approach) as well as the common leaf physiognomic methods (e.g., Climate Leaf Analysis Multivariate Program – CLAMP).

After extensive testing of CRACLE using modern floras the method was used to generate a >30,000 year record of climate for Western North America using the packrat (*Neotoma* spp.) midden plant macrofossils. This analysis provides a high resolution record of climatic change through the terminal Pleistocene glaciation and subsequent deglaciation. Many well-understood features of the global climate during this time based on independent data and models are captured in the packrat-CRACLE climate estimates including the Last Glacial Maximum, the Younger Dryas, the Holocene Thermal Optimum, and the late Holocene cooling trend.

CRACLE is presented here as an operational model for the reconstruction of climate from modern plant distributions and is shown to be applicable to Late Quaternary paleoclimate estimation from the plant fossil record. Application into deeper time is possible but will require further development of the method for estimating fossil taxon climate tolerance profiles; including both taxonomic and phylogenetic approaches.

BIOGRAPHICAL SKETCH

Robert S. Harbert has been working towards a Ph.D. in Plant Biology at Cornell University since 2011 studying the interaction of climate and plant distributions with goals of using vegetation to quantitatively estimate climate and to study the evolution of climate tolerances in the laboratory group of Dr. Kevin C. Nixon. He will graduate in August, 2016 with a Ph.D. in Plant Biology, and a concentration in Plant Systematics.

Before his work at Cornell he earned a Bachelors of Science in Biology at Roanoke College in Salem, VA, working on independent research under the guidance of Dr. Len Pysh. Starting in July, 2016, Robert will be joining the American Museum of Natural History for a 2 year appointment as a Gerstner Scholar in Bioinformatics and Computational Biology where he will continue his biodiversity, evolution, and climate research interests.

Robert is a member of the Botanical Society of America and has an active publishing record in their journal: *American Journal of Botany*.

Outside of academics, Robert is grateful for the support of his wife Lucy A.C. Harbert, and son Luke C. Harbert, and the rest of his family.

a) Professional Preparation

Roanoke College, Salem, VA	2011	Biology, B.S.
Cornell University, Ithaca, NY	2016	Plant Biology, Ph.D.

b) Appointments

Gerstner Scholar in Bioinformatics and Computational Biology, American Museum of Natural History. New York, NY. 2016-2018.

c) Honors and Awards

Outstanding Teaching Assistant Award, 2014-2015, College of Agriculture and Life Sciences (CALS), Cornell University, Ithaca, NY 14850.

Phi Beta Kappa, 2011

Alpha Chi, 2010

Summer Undergraduate Research Fellowship, 2010. American Society of Plant Biologists.

d) Products and Publications

Harbert, R.S., and K.C. Nixon. 2016. Applications of a novel model (CRACLE) for the estimation of >30,000 years of paleoclimate using packrat (*Neotoma* spp.) midden plant macrofossils from

the American Southwest. *33Rd Northeast-Midcontinent Paleobotanical Colloquium,* Cornell University, Ithaca, NY May 13-15, 2016.

Martinez, C., T.Y.S. Choo, D. Allevato, K. Nixon, W. Crepet, R. Harbert, C. Daghlian. 2016. *Rariglanda jerseyensis* a new ericalean fossil flower from the Late Cretaceous of New Jersey. *Botany* (in press)

Harbert, R.S., and K.C. Nixon. 2015. Climate reconstruction analysis using coexistence likelihood estimation (CRACLE): A method for the estimation of climate using vegetation *American Journal of Botany,* doi:10.3732/ajb.1400500

Harbert, R.S., A.H.D. Brown, and J. Doyle. 2014. Climate Niche Modeling in the Perennial *Glycine* (Leguminosae) Allopolyploid Complex. *American Journal of Botany* 101(4):710-721.

Harbert, R.S., and J. Doyle. Climate niche, invasiveness, and allopolyploidy: The case of perennial *Glycine* (Leguminosae). Presentation, Botany 2013, New Orleans, LA

Harbert, R.S. Growth and Nutrient Accumulation Responses to Phosphorus Deficiency in Cellulose Synthase Mutants of Arabidopsis thaliana. Poster presentation, 2011 Meeting of the American Society of Plant Biologists, Minneapolis, MN

Harbert, R.S. Root architecture responses to phosphorus in cellulose synthase mutants of *Arabidopsis thaliana*. Poster presentation, 2010 Meeting of the American Society of Plant Biologists, Montreal, ON, Canada

f) Teaching activities

Cornell University Graduate Teaching Assistant for:

BioG 1445 Introduction to Comparative Physiology (FA 2015, SP 2016)

BioPL 6410 Laboratory in Plant Molecular Biology (1wk Module FA, 2015)

BioPL 2420 Plant Function and Growth (SP 2013, SP 2015)

BioPL/Hort 2430 Cultivated Plant Taxonomy (FA 2014)

BioPL 2490 Hollywood Biology: Science and Cinema (SP 2014)

BioPL 2470 Economic Botany: Plants and People (FA 2013)

BioPL 2410 Introductory Plant Biodiversity and Evolution (FA 2012)

g) Collaborators and other affiliations

Kevin C. Nixon (Cornell University)

Jeffrey J. Doyle (Cornell University)

Anthony H.D. Brown (Centre for Australian National Biodiversity Research)

Camila Martinez (Cornell University)

Thereis Y.S. Choo (Cornell University)

Daniella Allevato (Cornell University)

Charles Daghlian (Dartmouth College)

ACKNOWLEDGMENTS

The work associated with this thesis would not have been possible without the extensive guidance and mentoring done by Dr. Kevin C. Nixon. Academic support was also found with Dr. William Crepet, Dr. Jeffrey J. Doyle Dr. Maria Alejandra Gandolfo, and Dr. Melissa Luckow. I also thank Thereis Choo and Daniella Allevato for their peer review of this and other manuscripts throughout my graduate career.

TABLE OF CONTENTS

PREFACE

The association of climate with plant distributions is intuitive at the coarsest levels. If shown images of a tropical forest and a typical desert most people (scientists or not) would be able to characterize the climate of the desert as “hot and dry” and that of the tropical forest as “warm and wet”, or something close. This is because the relationship of plants with their environment is observed everywhere we go. In the cold of winter, we see that plants have adapted to go dormant or otherwise avoid that seasonal extreme. In the heat of summer, the parched grass in the lawn indicates to the casual observer the effects of drought on plants. To the trained eye, physical characteristics of a natural vegetation, correlates of the characters of the constituent species, may provide insights into the precise nature of climate.

Individual plant species distributions, and the resulting vegetation types, have long been understood to be correlated with environmental factors, of which climate is a prevailing force (Gleason, 1926, 1939; Holdridge, 1947; Whittaker, 1956, 1967; Walter, 1973). Fundamental to the research and model development presented in this dissertation are the ideas laid out by Gleason's “Individual Concept of Plant Associations” (1926, 1939) and

Whittaker's (1956) “Vegetation of the Great Smoky Mountains” discussion of vegetation arrangement along environmental gradients.

The Climate Reconstruction Analysis using Coexistence Likelihood Estimation is the latest in a class of taxonomic methods for the estimation of climate from vegetation. The first in this class was the widely used Coexistence Approach (Mosbrugger and Utescher, 1997), followed closely by the lesser-used and nearly identical Mutual Climatic Range technique (Sinka and Atkinson, 1999). These methods have been widely applied to the fossil record (e.g., http://www.neclime.de/), but have also been criticized for being a relatively poor predictive method (Grimm and Denk, 2012; Grimm and Potts, 2015). CRACLE uses high quality point-locality data from specimen collections from a wide array of herbaria and other institutions distributed via the Global Biodiversity Information Facility (GBIF; www.gbif.org) and high resolution climate models (www.worldclim.org) to generate reliable and reproducible climate tolerance profiles for each taxon. Indirect comparison of CRACLE with these earlier methods showing the higher accuracy of CRACLE is illustrated in Chapter One.

As a paleoclimate proxy method, CRACLE is currently limited to parts of the fossil record for which plant fossils can be placed in

modern groups. CRACLE requires identifying extant, related analog taxa to use to characterize the climate tolerance of each fossil taxon. This is related to the Nearest Living Relative approach (Mosbrugger and Utescher, 1997). For convenience higher taxa (e.g., genera) may be substituted to represent a fossil taxon. Chapter Two tests how CRACLE responds to the use of genera to estimate modern climates from vegetation surveys. There is no other known test of this class of models with regard to the use of higher taxa despite it being often assumed when applying the Coexistence Approach and related methods to fossil floras (e.g., Boyle et al. 2008).

Given the potential issues with deep time applications the logical place to start with CRACLE as a paleoclimate proxy method is the Pleistocene. One of the most complete botanical fossil records of this time is for packrat (*Neotoma* spp.) macrofossils of the American Southwest. Curated by the United States Geological Survey (geochange.er.usgs.gov/midden/) this database includes more than 2500 fossil samples from the arid western U.S., and northern Mexico. For many of these samples the plant macrofossils have been identified. Given the relatively young age of most of these fossils the identifications could be made to a modern, extant genus or species. Chapter Three illustrates the results of the

application of CRACLE to a subset of the packrat data set in western North America. The primary product of this section is a >30,000 year time-line of climate change for this region that is consistent with global climate trends (e.g., the terminal Pleistocene glaciation then deglaciation) and exhibits unique features of climate change due to the ability of CRACLE to estimate a wide range of climate variables.

Moving forward from this point there are many methodological additions, adaptations, and changes that could improve CRACLE model performance. The field of geographic species distribution modeling (SDM) is a well-studied area of modern ecology and offers many methodological choices for the inference of continuous geographic distributions from presence-only point locality data. These methods attempt to extrapolate from presence localities to generalize across a study area to identify other areas of suitable environments. CRACLE, as currently described, is sensitive to geographic sampling biases and under-sampling so one avenue to test is the use of SDMs on the front-end of CRACLE to provide climate tolerance profiles that are more robust to these issues. Though unrelated to CRACLE, Chapter Four presents an application of SDM via the Maxent model (Phillips et al., 2004; Phillips and Dudik, 2008; Elith et al., 2011) to the

characterization of climate niche evolution in the *Glycine* subg. *Glycine* allopolyploid complex in Australia. Tools developed during this work to generate high-quality potential distribution models from presence-only data will be relevant to the linking of SDM to CRACLE as this research moves forward.

CHAPTER 1

CLIMATE RECONSTRUCTION ANALYSIS USING COEXISTENCE LIKELIHOOD ESTIMATION (CRACLE): A METHOD FOR THE ESTIMATION OF CLIMATE USING VEGETATION.

Previously published as:

Harbert, R.S., and K.C. Nixon. 2015. Climate reconstruction analysis using coexistence likelihood estimation (CRACLE): A method for the estimation of climate using vegetation. *American Journal of Botany,* doi:10.3732/ajb.1400500

Reprinted here with permission from the American Journal of Botany. Pagination here differs from the original.

Abstract

PREMISE OF THE STUDY: Plant distributions have long been understood to be correlated with the environmental conditions to which species are adapted. Climate is one of the major components driving species distributions. Therefore, it is expected that the plants coexisting in a community are reflective of the local

environment, particularly climate.

METHODS: Presented here is a method for the estimation of climate from local plant species coexistence data. The method, Climate Reconstruction Analysis using Coexistence Likelihood Estimation (CRACLE), is a likelihood-based method that employs specimen collection data at a global scale for the inference of species climate tolerance. CRACLE calculates the maximum joint likelihood of coexistence given individual species climate tolerance characterization to estimate the expected climate.

KEY RESULTS: Plant distribution data for more than 4000 species were used to show that this method accurately infers expected climate profiles for 165 sites with diverse climatic conditions. Estimates differ from the WorldClim global climate model by less than 1.5°C on average for mean annual temperature and less than ~250 mm for mean annual precipitation. This is a significant improvement upon other plant-based climate-proxy methods.

CONCLUSIONS: CRACLE validates long hypothesized interactions between climate and local associations of plant species. Furthermore, CRACLE successfully estimates climate that is

consistent with the widely used WorldClim model and therefore may be applied to the quantitative estimation of paleoclimate in future studies.

Introduction

It has long been documented that plant species distributions, vegetation associations, biodiversity, and plant morphological traits are correlated with climate (see Bailey and Sinnott, 1915, 1916; Gleason, 1926; Holdridge, 1947; Whittaker, 1956, 1967; Walter, 1973; Diaz et al., 1998). The varying ability of plants to tolerate climatic extremes is a major factor contributing to the global range of taxonomic and physical diversity observed in plant communities. The presence of plant species within standing vegetation is thus reflective, and potentially predictive, of long-term climatic conditions and trends.

Local climate estimation via methods that use vegetation data potentially provide a characterization of the climate that led to the formation of that vegetation. Variation in vegetation composition occurs at a very fine resolution, especially in regions with high geodiversity (e.g., elevation changes over short distances). A set of approaches that we term Estimation of Climate from Vegetation (ECV) has been used widely in paleobotanical studies that apply leaf

morphological characters or taxonomic co-occurrence within a community (modern or fossil) to infer past climates (e.g., Prentice et al., 1991; Herman and Spicer, 1996; Herman and Spicer, 1997; Mosbrugger and Utescher, 1997; Sinka and Atkinson, 1999; Kennedy et al., 2002; Kühl et al., 2002; Sharpe, 2002; Kowalski and Dilcher, 2003; Kühl and Litt, 2003; Roth-Nebelsick et al., 2004; Yang et al., 2007; Punyasena, 2008; Thompson et al., 2008, 2012; Velasco-de León et al., 2010; Srivastava et al., 2012).

Here, we present a new approach to ECV that leverages publicly available distribution data based explicitly on specimen collections (obtained via the Global Biodiversity Information Facility [GBIF]) to estimate local climate parameters using the species composition of the vegetation and the climate tolerance of those species in an empirical likelihood modeling framework. Leveraging the data publicly available via GBIF allows potential users of this method to easily generate species climate tolerance estimates and community climate estimates at a global scale. Unmodified, this method can be applied to recent paleoclimatic inference (e.g., Holocene or Pleistocene) using data from pollen core samples or pack rat midden macrofossils (e.g., LaMarche, 1973; Scuderi, 1987; Kühl et al., 2002; Sharpe, 2002; Kühl and Litt, 2003; Paciorek and McLachlan, 2009; Thompson et al., 2012). For the present

application, it is assumed that evolution of climate tolerance within species has been slow in this time period, an assumption consistent with other taxonomic ECV methods (e.g., Prentice et al., 1991; Mosbrugger and Utescher, 1997; Kühl et al., 2002). With consideration of phylogeny and changing niches, in the future the proposed method may be capable of quantitative climate inference in deeper time.

Taxonomic coexistence ECV— The term “coexistence” has been applied to an alternative set of ECV methods that use taxonomic (usually species) composition of the vegetation at a particular locality to estimate climate parameters (see, especially, Mosbrugger and Utescher, 1997; Punyasena, 2008; Thompson et al., 2012). In general, the Coexistence Approach (Mosbrugger and Utescher, 1997) and the Mutual Climatic Range method (Sinka and Atkinson, 1999; Sharpe, 2002; Thompson et al., 2012) estimate climate from taxonomic coexistence by calculating the overlap of species climate envelopes—multidimensional climate space where a species is known to occur. This methodology was originally proposed as a technique for inferring paleoclimate in association with the Nearest Living Relative method, which uses the climate tolerance of the extant sister taxon of a fossil as a surrogate climate

envelope (Mosbrugger and Utescher, 1997). The Coexistence Approach for the estimation of modern climate has been criticized as sensitive to error in the estimation of individual species' climate tolerance, resulting in poor performance and bias in quantitative estimates of climate from taxonomic coexistence data (Grimm and Denk, 2012).

A derivation of the Coexistence Approach was proposed using univariate likelihood models of climate tolerance at the family level (Kühl et al., 2002; Punyasena, 2008). For the models presented by Punyasena (2008), climate tolerance likelihood distributions were estimated from distributional data documented in the Gentry Forest Transect data set (Gentry, 1988). The narrow geographic scope of that study—fewer than 250 collection sites, mostly Neotropical—and the pooling of occurrence data by family led to poor performance, as evidenced by low, and often non-significant, statistical correlation between the model prediction of climate and the observed climate (Punyasena, 2008). The Probability Density Function method, described by Kühl et al. (2002) and implemented by Kühl and Litt (2003), made a proposal similar to, but preceding, Punyasena (2008). However, neither study (Kühl et al., 2002; Kühl and Litt, 2003) provided any validation of the method against known or estimated modern climate. Similar methods using

vegetation composition to estimate “floristic temperature” through various modeling strategies—for example, weighted averaging partial least squares, Breiman’s Random-Forest Regression (Bertrand et al., 2011), weighted (by abundance) average of species averages (De Frenne et al., 2013), and correlation between temperature and average Ellenberg Indicator Values (Lenoir et al., 2008)—have been proposed in the context of investigating vegetation “lag” behind modern climate warming in Europe. To date, no explicit validation of the performance of these models (Lenoir et al., 2008; Bertrand et al., 2011; De Frenne et al., 2013) in relation to actual or inferred climate records has been reported. Blonder et al. (2015) propose a method (provided in the R package “comclim”) for the identification and quantification of vegetation community climate lag that may be relevant to all taxonomic ECV methods in future studies.

Physiognomic ECV— The observation that climate is correlated with leaf physiognomy –variation of leaf morphology within vegetation—has a long history (Bailey and Sinnott, 1915, 1916) and has led to multiple related ECV methods (Wolfe, 1993, 1995; Wilf, 1997; Royer et al., 2005; Spicer et al., 2009; Peppe et al., 2011). Most often, aside from testing on modern vegetation,

physiognomic methods have been applied in paleobotanical studies (e.g., Herman and Spicer, 1996, 1997; Kennedy et al., 2002; Kowalski and Dilcher, 2003; Roth-Nebelsick et al., 2004; Yang et al., 2007; Velasco-de León et al., 2010; Srivastava et al., 2012). Physiognomic methods are usually founded on observed correlations between leaf margin variation (as percentage samples of the standing vegetation, independent of species composition) and climate (e.g., the Climate Leaf Analysis Multivariate Program [CLAMP]; Wolfe, 1993, 1995; Spicer et al., 2009). As such, physiognomic methods require intensive sampling within vegetation to establish the presence and relative abundance of each physiognomic character. CLAMP and other related methods have been criticized because correlations between leaf form, abundance in vegetation, and climate may be geographically variable (Greenwood et al., 2004, Jacques et al., 2011) and inconsistent across and between taxonomic groups (Little et al., 2010). Although such issues are important, and may reduce accuracy in some scenarios, physiognomic ECV methods have been highly successful and continue to be widely used.

Here, we propose a new ECV method called Climate Reconstruction Analysis using Coexistence Likelihood Estimation (CRACLE). Implementation of CRACLE requires three data sources:

(1) a verified species list for a site, (2) detailed georeferenced locality records characterizing the global distribution of each species occurring at the study site, and (3) a continuous, georeferenced global climate model that can be used to generate species climate profiles from the species distribution data. It is important to note that CRACLE does not require density, proportion, percent cover, or abundance sampling of either species demographics or physiognomic characters. Instead, CRACLE uses explicit species occurrence data for coexistent species and selects the most likely climate parameters for a given site, based on maximizing the univariate empirical joint likelihood of species' climatic coexistence. Here, we test the CRACLE method using species distribution data obtained through GBIF (http://www.gbif.org) and the WorldClim global climate models (Hijmans et al., 2005) to generate species–climate tolerance profiles and to evaluate the performance of CRACLE relative to WorldClim.

Materials and Methods

Conceptual methodology— CRACLE is based on the principle that the coexistence of plant species at a locality implies that the climate is suitable, to varying degrees, for all species present (Gleason, 1926). It is therefore closely related to other methods,

including the Coexistence Approach (Mosbrugger and Utescher, 1997), the Probability Density Function method (Kühl et al., 2002), and the Mutual Climatic Range method (Sinka and Atkinson, 1999; Thompson et al., 2012). CRACLE hypothesizes that the probability of the occurrence of a single species Sp_i given a specific value θ of a climate variable can be described, assuming the existence of a probability function of occurrence in relation to that climate variable. In the present study, two types of distributions are used to calculate this probability: a normal distribution for a parametric likelihood function (P-CRACLE) and a Gaussian kernel density estimation as a nonparametric likelihood (Silverman, 1986) function (N-CRACLE). The likelihood that a value for any climate variable is the same as that of the target site, given the joint likelihood of finding all the coexisting species ($Sp_1 - Sp_i$) in that climate (θ), must then be the maximum of the sum of the empirical coexistent species likelihood functions:

$$L(Sp_{1:i};\theta) = \sum_{n=1}^{i} \ln\left(Pr(Sp_n|\theta)\right)$$

A simulated example of this approach is provided in Fig. 1.1. Using a generalized likelihood function, climate variables are evaluated independently. Results are pooled to give the inferred climate profile for the site.

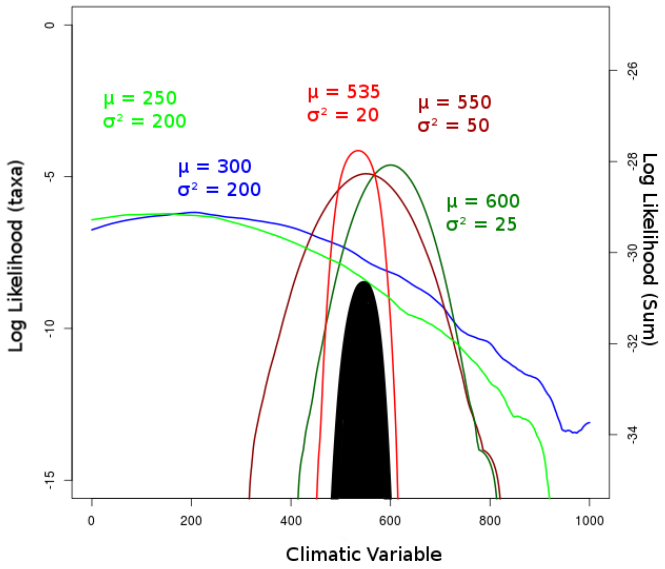

Figure 1.1 Likelihood curves and representation of the joint likelihood for five simulated species. Five univariate species-niche-occupancy curves were each generated by randomly sampling normal distributions with given hypothetical population means and standard deviations. The sum log-likelihood (joint likelihood) curve is represented by the dark gray shaded area and is referenced against the right y -axis. The optimum of the sum log-likelihood curve is the most likely climate value at which to find all five species co-occurring. Note that this optimum appears to be most influenced by the narrowly distributed (low standard deviation) hypothetical species curves.

The P-CRACLE method assumes that a species' occurrence is normally distributed along climatic variables, whereas the N-CRACLE method instead makes no assumption about the shape of the distribution in climatic space, other than that the distribution is continuous and smooth. It is not the goal of the present study to evaluate how well these models approximate individual dimensions of the realized climate niche of a species (Hutchinson, 1957;

Peterson et al., 2011). It is assumed that the sample of data available for each species is representative of the actual distribution and reflects patterns and ecological biases present at the species-level niche. The CRACLE program is publicly available as a simple R (R Core Team, 2014) script (Appendix S1.1).

Species distribution data— At the time of this study, GBIF has compiled >100 million georeferenced records on the distribution of plants. This immense repository of primary distribution data is essential for the CRACLE methodology to be streamlined and easily applied to many sites and between geographically distant areas. Georeferenced records for relevant taxa for each site (species lists: Appendix S1.2) were downloaded through the GBIF data portal (lists of primary data providers: Appendix S1.3). Semiautomated data processing then removed records that did not include exact latitude and longitude coordinates. Missing coordinates, integer coordinates (poor precision), and otherwise incorrect coordinates (e.g., occurring outside of the referenced country or in the wrong hemisphere) were common categories of error with GBIF data that resulted in records being excluded from our study. Species that were represented by fewer than five records were excluded as well, to avoid over-fitting of the likelihood functions due to under-

sampling of the distribution. The limit of five records, though arbitrary, was based on the observation that species with very few records in GBIF seemed to be biasing CRACLE via artificially narrow samples of climatic tolerance. Records from GBIF used in our study were contributed by 1158 unique data providers (Appendix S1.3). More than 3.75 million unique records for 4388 species (Appendix S1.2) were obtained from GBIF.

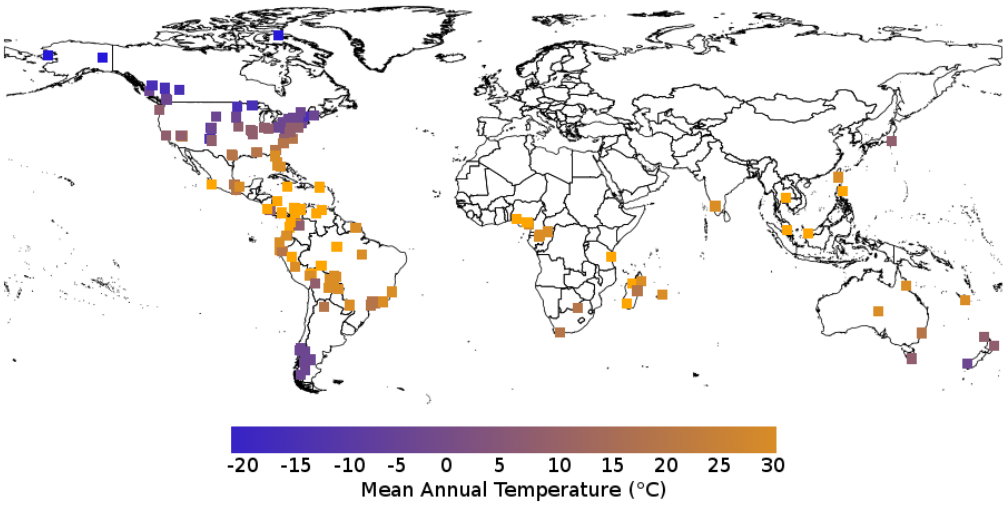

Figure 1.2 Geographic distribution of test sites. Estimates of mean annual temperature from WorldClim (Hijmans et al., 2005) have been used to color the study sites (for site data, including references and localities, see Supplemental Data with the online version of this article; Appendix S1.4).

Survey sites— A total of 165 sites (locations: Fig. 1.2; species lists: Appendix S1.2) were selected to represent a wide variety of climatic regimes. The 165 sites are located on five continents

(excluding Europe and Antarctica), occur between −50 ° S and 73 ° N, and are found from sea level to a maximum of 3500 m. Analysis of any sites in Europe was excluded from our study because of the abundance of cultivated plant records and nonrandom spatial bias across political boundaries present in GBIF data for that region (Yesson et al., 2007), which would bias the climate tolerance profiles and, therefore, the CRACLE projections. Four vegetation-survey data sets (Gentry, 1988; Boyle, 1996; Killeen and Schulenberg, 1998; Phillips and Miller, 2002; Webb et al., 2003; Jenkins and Motzkin, 2009) provided 107 of the study sites. The Gentry Forest Transect data set (Gentry, 1988; Phillips and Miller, 2002), Boyle's Neotropical surveys (Boyle, 1996), and the Noel Kempff Forest data set (Killeen and Schulenberg, 1998) are direct surveys of woody vegetation primarily from the Neotropical forests. Each of these data sets is available for public use from the Synthesis and Analysis of Local Vegetation Inventories Across Scales Project (SALVIAS, http://www.salvias.net). One site for analysis was developed from the Harvard Forest 1980–2009 database by trimming observed taxa to include only woody species (Jenkins and Motzkin, 2009). Tree plot surveys of the vegetation of the Nevada nuclear test site (Webb et al., 2003) were also included. Various other sites were developed using national, provincial, and

state park floras and checklists in the United States, Canada, South Africa, New Zealand, and Australia (see Appendix S1.4). Survey data for Argentina and Chile were compiled during a course for undergraduates at Cornell University by K.C.N and students in January 2014 (Appendix S1.2).

Species lists for each site are reported in Appendix S1.2. These species lists are not necessarily as complete as those reported in the sources listed. Specifically, the lists were limited to trees and shrubs only where those are the dominant life forms. This is based on the assumption that woody plants are generally more sensitive to the annual climate and therefore have narrower climate niches (Smith and Beaulieu, 2009), compared with grasses and annuals, because of year-round exposure. However, the CRACLE method does not specifically require this restriction. Species lists were trimmed further if identification was made only to genus or if the species names were not found in GBIF and/or if insufficient distributional data were available from GBIF (<5 records).

Climate model data— Climate data are from the downscaled 2.5-arcminute resolution (~0.041667 degrees) WorldClim model. The WorldClim model is a high-resolution continuous grid of interpolated climate data for the world's land areas derived from

>40 000 weather stations around the world. The original resolution of 0.5 arcminutes was downscaled to 2.5, 5, and 10 arcminutes, of which we chose to use the 2.5 arcminute grid to better match the potential for spatial uncertainty in the majority of the georeferenced GBIF data (Guralnick et al., 2006; Hill et al., 2009).

The WorldClim grid was chosen over other interpolated climate models for its relatively higher resolution (2.5 arcminutes vs. 30 arcminutes in New et al., 2002) and its global, rather than North American, extent (Thornton et al., 1997; Daly et al., 2000, 2002, 2008). WorldClim also includes the largest network of weather station data (>40 000 stations used to develop the model for temperature) among these gridded models (Hijmans et al., 2005). It is important to note that any interpolated climate model for the Earth's land surfaces would be suitable for the basis of and/or evaluation of CRACLE and that the relative performance of CRACLE in the context of all available models should be the focus of future research.

We focused on the estimation of temperature and selected precipitation variables from vegetation data. Specifically, we estimated (1) mean annual temperature, (2) average annual maximum temperature, (3) average annual minimum temperature, (4) mean annual precipitation, (5) precipitation of the three

consecutive wettest months, and (6) precipitation of the three consecutive driest months. This coincides well with the inferences put forth by other methods of ECV. CRACLE can potentially be applied to any environmental variable but has not yet been tested beyond these six climate variables.

Climate tolerance profiles— Plant distribution data were compiled from GBIF for each species. The species climate tolerance profile, or "niche," is simply the set of climate parameters in which a species is reported to occur, based only on the locality data referenced to the climate model (following the logic of the "realized niche"; Hutchinson, 1957; Peterson et al., 2011). Climate values are extracted from the WorldClim database (Hijmans et al., 2005) for each occurrence locality of a species. All calculations were performed using R (R Core Team, 2014) and the R software package "raster" (Hijmans, 2014).

Evaluation methods— The performance of CRACLE in relation to the WorldClim models was evaluated by calculating the Pearson's correlation of CRACLE inference for each locality with WorldClim estimates, the average difference of the median CRACLE values in relation to WorldClim, the average minimum difference between

CRACLE and WorldClim, and the mean prediction range. The average difference is calculated as the average of the absolute difference between the median of each CRACLE interval and the WorldClim estimate. The average minimum difference is the average of the absolute minimum difference between the CRACLE interval and the WorldClim estimate for each site. To clarify: The minimum difference between the interval 1:5 and the value of 4 is zero, whereas the median of the interval 1:5 is 3 and has a difference of 1 from the value 4. Both values measure the accuracy of CRACLE in relation to WorldClim, but the average difference considers the entire CRACLE interval whereas the average minimum difference considers only the shortest distance between the CRACLE interval and the WorldClim value. The mean predicted range (MPR) is a measure of precision. This value is calculated as the absolute value of each predicted range. Smaller MPRs indicate that the optimal climate values are narrower under which the species observed at a site could coexist, given the likelihood criteria (P-CRACLE or N-CRACLE) used.

To assess the effect of the sample of taxa in relation to the entire vegetation on the output of CRACLE, two sites were targeted for a simulation of incomplete taxonomic representation. The Harvard Forest and Barro Colorado Island data sets (including 165

and 99 species, respectively) were chosen for this analysis because of the abundance of species data and to represent both tropical and temperate regions. For each, random partitions of the species were made, without replacement, at intervals of 3% from 3–99% of the total population of species. At each interval, partitions were drawn 100 times and CRACLE was performed. The mean and standard deviation of the optimal estimate of each climate variable at each sampling interval were calculated. To more generally assess the effect of sample sizes on performance, the outputs of CRACLE for all 165 sites were used to calculate the Pearson's correlation between CRACLE errors in relation to WorldClim versus the number of species observed at a site and the number of GBIF records used in CRACLE. These statistics summarize the relative effects of sample size and data availability on the success and relative stability of CRACLE compared with WorldClim.

The resolution of an interpolated climate model like WordClim determines the extent to which fine-scale climatic variation can be represented by the models. Here, the 2.5-arcminute WorldClim grids are used to best match the confidence with which the average georeferenced specimen record can be placed (Guralnick et al., 2006; Hill et al., 2009). However, a higher-resolution (0.5 arcminutes) version of WorldClim is available that has been shown

to represent greater local climatic variation than the lower-resolution (2.5 and 10 arcminutes) version (Hijmans et al., 2005). Locality data for each site were used to resample the higher-resolution grid values within 1 arcminute in any direction from the cell in which the site was located, resulting in a summary of climatic variation within a 2.5-arcminute square around each site. Given the potential for georeferencing errors in the site coordinates, this procedure considers local climate uncertainty rather than the absolute values provided by the 2.5-arcminute WorldClim grids. Whether the range of local climate extracted from the 0.5-arcminute grid overlapped with CRACLE for each site was assessed, as was the overlap of CRACLE with the values extracted from the 2.5-arcminute grids for comparison.

Results

Climate estimates from WorldClim for each of the 165 study sites (Fig. 1.2 and Appendix S1.4) showed mean annual temperatures from -16°C to 28°C, with an average of 16.3°C (Appendix S1.4). Estimated (Hijmans et al., 2005) average annual precipitation for the study sites also showed extreme variation, with a range of 136–7407 mm and a mean of 1483 mm (Appendix S1.4).

Reconstructions of mean annual, maximum annual, and

minimum annual temperature for each site using CRACLE (Fig. 1.3 and Appendix S1.4) are highly consistent with the WorldClim estimated values for these sites. The P-CRACLE method yielded an average absolute difference of 1.4°C for the inference of mean annual temperature, 1.4°C for maximum annual temperature, and 2.1°C for minimum annual temperature (Fig. 1.3); whereas the N-CRACLE method yielded an average absolute difference of 1.3°C for mean annual temperature, 1.2°C for maximum annual temperature, and 1.6°C for minimum annual temperature. Both methods return results that exhibit strong linear correlation with the WorldClim values for all temperature variables ($\rho \geq 0.94$; Fig. 1.3). CRACLE inference of mean annual precipitation, precipitation of the three consecutive wettest months, and precipitation of the three consecutive driest months show similar overall performance for the N-CRACLE (average error: 56−169 mm, $0.96 > \rho > 0.6$; Fig. 1.3) and P-CRACLE results (average absolute difference: 50−251 mm, $0.91 > \rho > 0.75$; Fig. 1.4). Note, however, that WorldClim estimates only single values for each cell in the model grid and, therefore, these estimates of error are conservative, given that there is inherent uncertainty in the WorldClim models.

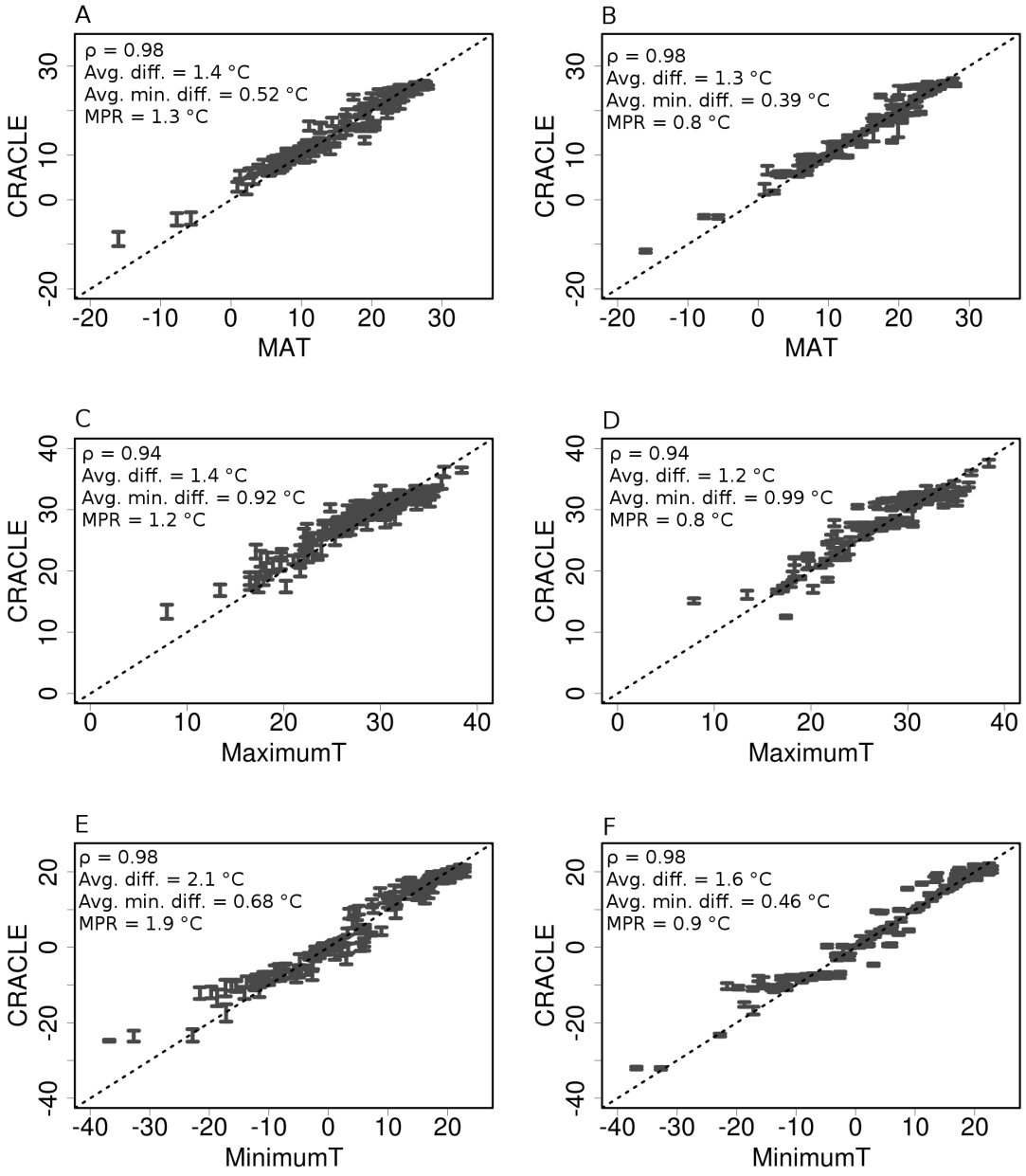

Figure 1.3 Temperature results: Plots of the WorldClim modeled climate values (Hijmans et al., 2005) versus the CRACLE results. For mean annual temperature (A, B), maximum temperature (C, D), and minimum temperature (E, F). The P-CRACLE (A, C, E) and N-CRACLE (B, D, F) results are shown. Temperature values are in°C. Pearson's correlation (ρ), the average difference of the median value for CRACLE in relation to WorldClim, the average minimum difference between CRACLE and

WorldClim, and the mean prediction range (MPR) are reported in each plot.

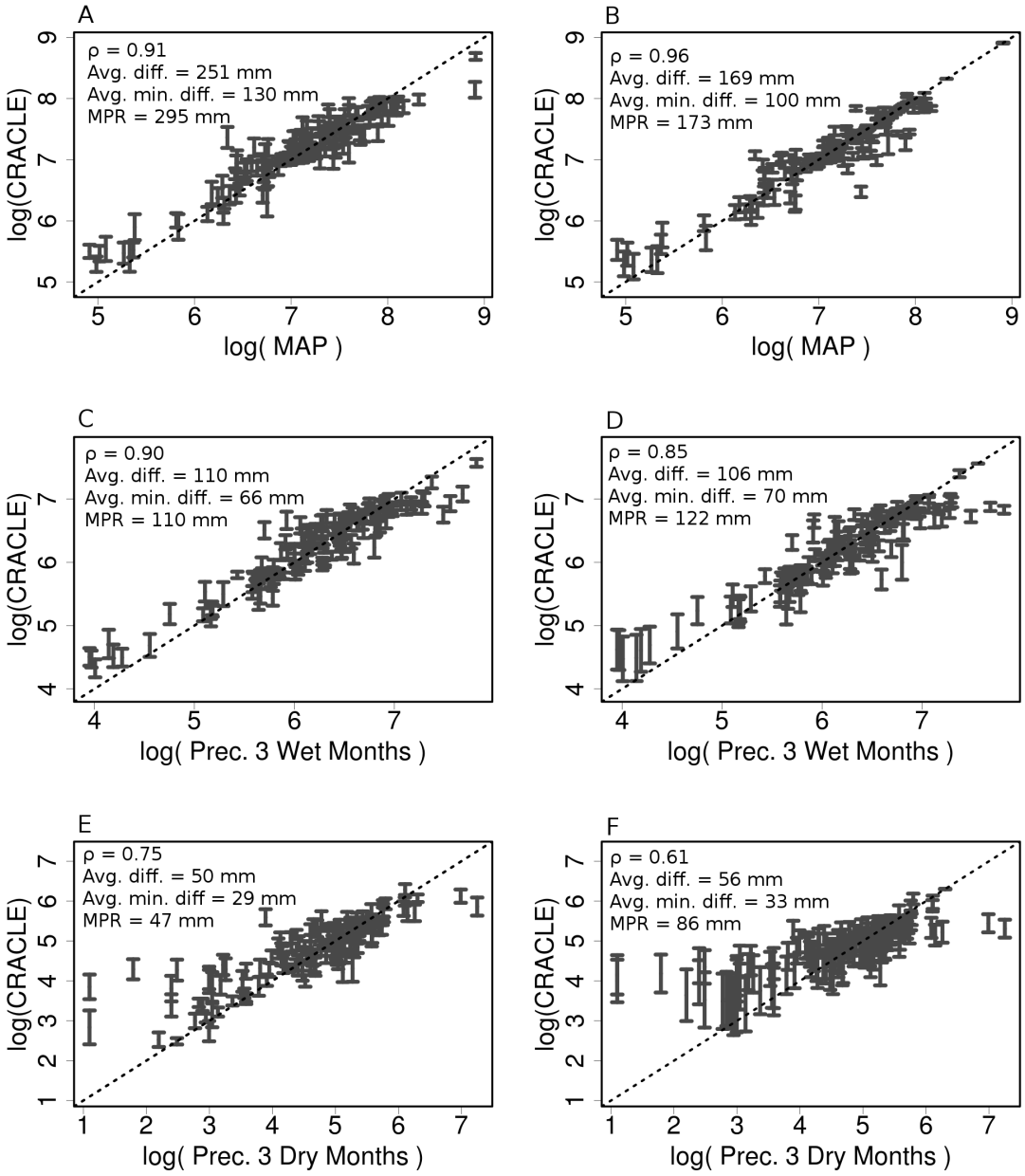

Figure 1.4 Precipitation results: Plots of the WorldClim modeled climate values (Hijmans et al., 2005) versus the CRACLE results. For mean annual precipitation (A, B), precipitation in the 3 consecutive wettest months (C, D), and

precipitation in the 3 consecutive driest months (E, F). The P-CRACLE (A, C, E) and N-CRACLE (B, D, F) results are shown. Precipitation values are millimeters plotted on a log scale. Pearson's correlation (ρ), the average difference of the median value for CRACLE relative to WorldClim, the average minimum difference between CRACLE and WorldClim, and the mean prediction range (MPR) are reported in each plot.

No significant correlation between the species sample size and the performance of CRACLE in relation to WorldClim was observed (Table 1.1). Furthermore, the number of GBIF records included in a CRACLE analysis showed no correlation with the relative CRACLE performance (Table 1.2). The resampling performed for the Harvard Forest and Barro Colorado Island sites indicates that a rapid stabilization of CRACLE was achieved with relatively small samples of taxa (Fig. 1.5).

Table 1.1: Pearson's Correlation Coefficients of CRACLE relative differences from WorldClim vs. the number of species observed at each site.

	P-CRACLE	N-CRACLE
Mean annual temperature	-0.24	-0.09
Minimum annual temperature	-0.24	-0.00
Maximum annual temperature	-0.18	-0.05
Mean precipitation	0.11	-0.03
Precipitation of the 3 wettest months	0.05	0.01
Precipitation of the 3 driest months	0.15	0.13

Resampling the 0.5-arcminute WorldClim model for

hypothetical climatic variation in the vicinity of each site analyzed with CRACLE showed an overall increase in the accuracy of CRACLE in comparison to WorldClim. Taking into account local climatic variation is one way to allow for uncertainty in the geographic position of each vegetation survey site as well as the inherent uncertainty present in the WorldClim models. The percentage of sites where CRACLE and WorldClim intersect (yielding a “correct” inference) increased for both CRACLE models and for all variables used (Fig. 1.6).

Discussion

Comparison to taxonomic ECV methods — Implementation of the Coexistence Approach method (Mosbrugger and Utescher, 1997) has been shown to return overly broad (mean prediction range ~5°C) estimates of mean annual temperature (Grimm and Denk, 2012). In addition, convergence of mean annual temperature estimates on a single range (Grimm and Denk, 2012), and low linear correlation (low accuracy) between inferred and expected climate variable values (Punyasena, 2008) are seen in taxonomic ECV methods. The Mutual Climatic Range method (Sinka and Atkinson, 1999) as implemented by Thompson et al. (2012) with a percentile weighting strategy is an improvement over the Coexistence

Approach or the unweighted Mutual Climatic Range method. Weighted Mutual Climatic Range estimates of climate across North America showed narrower average ranges and good accuracy, with a lower average error (2.9°C) and high correlation (0.97) for the estimation of mean temperature of the coldest month (Thompson et al., 2012).

By contrast, both CRACLE methods return more accurate (mean error: P-CRACLE, 1.4°C; N-CRACLE, 1.3°C), narrower (MPR: P-CRACLE, 1.2°C; N-CRACLE, 0.8°C), and more distinct (Fig. 1.3A, B) estimates for mean annual temperature. Similar results are obtained for the other climate variables estimated by CRACLE (Figs. 1.3C–F and 1.4). Compared to the method presented by Punyasena (2008), CRACLE demonstrates improved statistical correlation between climate inference and site data (Tables 1.3, 1.4, and 1.5). Results for both CRACLE methods reject previous conclusions that taxonomic methods of ECV are capable only of qualitative estimation of climate (Grimm and Denk, 2012).

CRACLE sensitivity to sampling — We also tested the sensitivity of CRACLE to the number of species observed at a locality. No correlation was observed between the number of species at a site and the performance of CRACLE compared with

WorldClim, presumably because all the test sites had sufficient species samples for CRACLE (Table 1.1). However, even though there is no broad effect, it is expected that there will be a within-site effect, because adding species to the list should lead to narrower and narrower estimates. To test this effect and develop an idea of how many species may be required for an accurate CRACLE estimate, a resampling procedure was carried out for the Harvard Forest and Barro Colorado Island sites that takes random partitions of the total species list. These results suggest that as more taxa are added to a CRACLE analysis, the stability of the result increases, regardless of method or variable (Fig. 1.5). However, N-CRACLE appears to stabilize more rapidly than P-CRACLE (Fig. 1.5), which suggests that stable answers may be obtained with fewer taxa from N-CRACLE than from P-CRACLE. However, the number of taxa required for a stable result seems to vary between the two sites and to depend on which climate variable is being analyzed (Fig. 1.5). It may be that more taxa are required for a stable result for temperature using N-CRACLE in temperate regions versus lowland tropical regions, but the opposite may be true for the estimation of precipitation.

Table 1.2: Pearson's Correlation Coefficients of CRACLE relative differences from WorldClim vs. the number of individual collection record localities used.

	P-CRACLE	N-CRACLE
Mean annual temperature	-0.12	0.01
Minimum annual temperature	-0.12	0.02
Maximum annual temperature	-0.08	0.02
Mean precipitation	0.02	-0.00
Precipitation of the 3 wettest months	-0.04	-0.04
Precipitation of the 3 driest months	0.07	0.05

CRACLE fine-scale climatic detection — We expect that the sensitivity of vegetation to microclimate heterogeneity enables CRACLE to estimate climate on a scale that is much finer than the WorldClim grids (~4 km square at the equator). Consequently, although the CRACLE anomalies (and inferred performance) are based on differences from WorldClim estimates, it is likely that at least some of this difference is due to this sensitivity of vegetation and the precision of CRACLE. The finer-resolution (0.5-arcminute) WorldClim grids were sampled to observe the hypothetical local climatic variation proposed by the WorldClim model at that resolution, but not in the 2.5-arcminute grids used up to this point. Note that the 0.5-arcminute grid is not used for the initial stages of CRACLE because of the spatial uncertainty present in the point-

coordinate species-distribution data used in the present study (Guralnick et al., 2006; Hill et al., 2009). The result of this analysis showed that when finer local climatic variation based on the WorldClim grid was taken into account, the number of sites where CRACLE and WorldClim intersect increases for both CRACLE models (Fig. 1.6). This supports the idea that CRACLE may be representing climatic variation occurring at a higher resolution than 2.5 arcminutes. However, without independent validation of the climate at these or other vegetation sites, it is not possible to say that CRACLE is performing quantitatively better at the native WorldClim resolutions or finer.

Comparison to physiognomic ECV methods — To compare general performance of CRACLE with available physiognomic methods (Royer et al., 2005; Spicer et al., 2009; Jacques et al., 2011; Peppe et al., 2011), we generated the same performance statistic s as for the CRACLE methods—mean error, Pearson's correlation, and Spearman's correlation—from published results for those methods (Royer et al., 2005; Jacques et al., 2011; http://clamp.ibcas.ac.cn/ PhysgAsia1_Files/ResAsia1.xls). The MPR produced by these methods is zero because only single values, not ranges, are estimated and therefore the precision cannot be

compared with CRACLE. However, CRACLE results show more accurate inference of climate than any of these physiognomic ECV methods when results are compared across different sets of modern sites (Figs. 1.3 and 1.4; Tables 1.3, 1.4, and 1.5). For example, CRACLE estimates of mean annual temperature yielded an average error of 1.4°C (P-CRACLE) and 1.3°C (N-CRACLE), compared with 1.8°C for CLAMP (Jacques et al., 2011) and ~2°C for leaf margin analysis (Royer et al., 2005).

As a paloeclimate estimation method — The potential application of CRACLE for paleoclimate estimation will initially be limited to the late Quaternary (e.g., Holocene or Pleistocene) because of the need to use extant species distributions in calculations. The CRACLE method can be applied to pollen microfossils or pack rat midden debris where taxonomic placement of fossils can be made in relation to modern species (e.g., LaMarche, 1973; Scuderi, 1987; Kühl et al., 2002; Kühl and Litt, 2003; Thompson et al., 2008; Paciorek and McLachlan, 2009). For CRACLE to be applied in these cases, conservation of the realized climate niche for each species must also be assumed to be constant, or at least nearly so, and vegetation composition must track long-term climate change as a result (e.g., Prentice et al., 1991;

Martínez-Meyer and Peterson, 2006; Pearman et al., 2008; Losos, 2008; Couvreur et al., 2011). Using the pollen record of the Late Quaternary, Veloz et al. (2012) observed that for some North American taxa, realized climate niches have shifted, while for other taxa realized climate niches appear to be stable. In some cases, the shift in realized niche was from a climate with no modern analog—a case where the modern distribution of a taxon will be a particularly poor estimation of past realized niche. How-ever, as long as shift s in realized niches are not strongly correlated across fossil taxa and between climate variables, CRACLE should continue to perform well. CRACLE also has the potential to identify non-analog climates because the likelihood function for each variable is estimated independently.

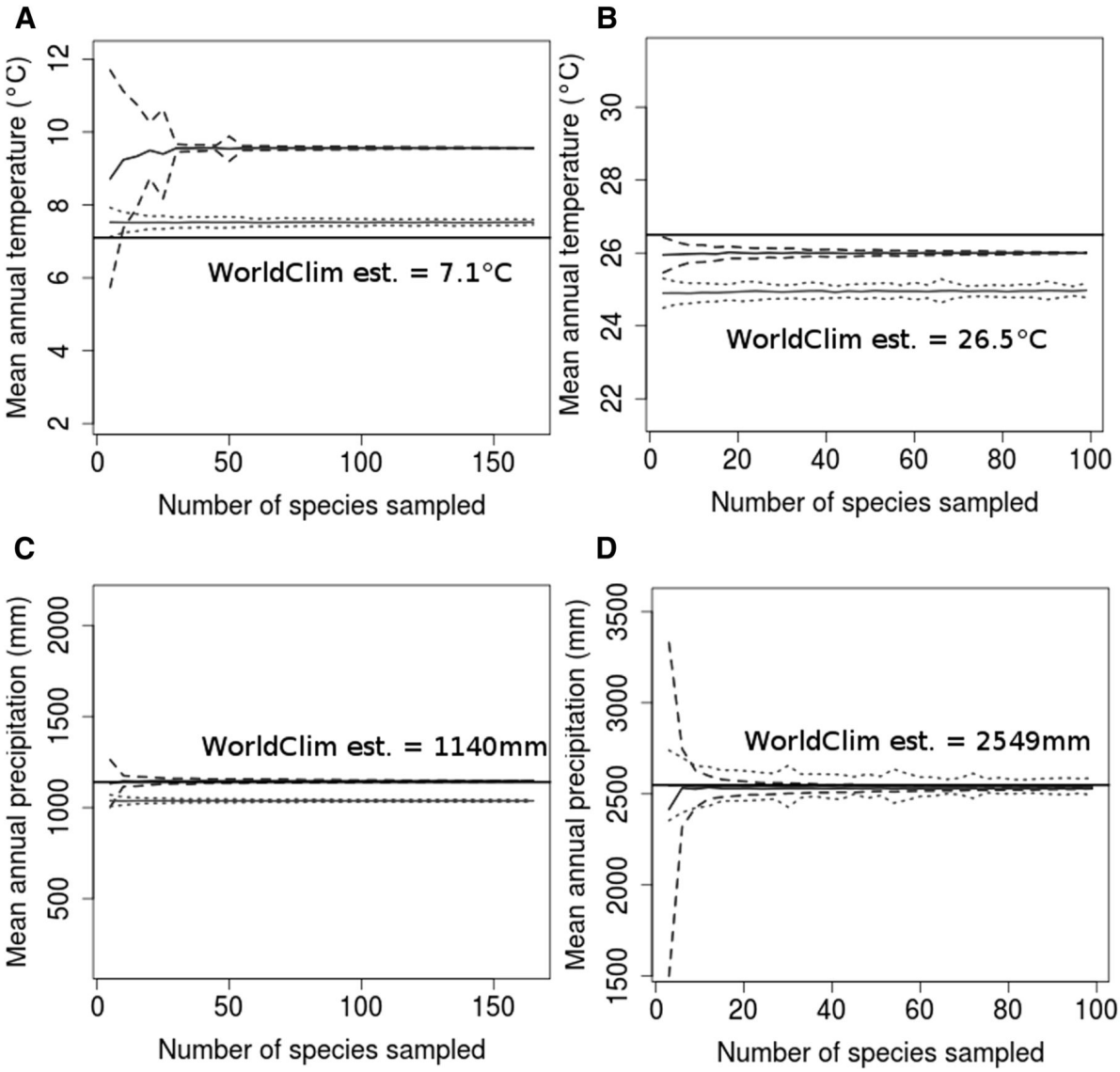

Figure 1.5. Resampled CRACLE for the Harvard Forest and Barro Colorado Island sites to examine the effect of sample size on stability of prediction. Random samples of taxa were taken from the species lists for the Harvard Forest and Barro Colorado Island sites of increasing size from 3% to 99% (by increments of 3%). At each sample size, 100 repetitions were made and the P-CRACLE and N-CRACLE optima were recorded. The plots here show the mean of the optima (solid lines) and the 95% confidence interval (dashed lines) for both the N-CRACLE (dotted) and P-CRACLE (dashed) results. Mean annual temperature (A,B) and mean annual precipitation (C, D) are reported. The solid horizontal line represents the WorldClim (WC) estimate for each

site and variable.

Relative benefits of a streamlined taxonomic ECV method – Physiognomic ECV methods have been applied to climate reconstruction at much deeper time (e.g., Miocene: Yang et al., 2007) due to the hypothesized constancy of the correlations between climate adaptation and morphology. However, CRACLE may be preferable to physiognomic methods for analysis using modern vegetation because of the easily collected data—species lists—as opposed to detailed measurements of leaf samples necessary for physiognomic methods. As seen in this study site, the data required for CRACLE are easily harvested from existing sources or in the field; various species lists (Appendix S1.2) compiled by directed vegetation surveys (Gentry, 1988;Boyle, 1996; Killeen and Schulenberg, 1998; Phillips and Miller, 2002; Webb et al., 2003), national and state park monitoring (Appendix S1.4), long-term ecological studies (Jenkins and Motzkin, 2009), and academic field-trip studies can all be successfully used as input for CRACLE. It is also important to note that a complete sample of species occurring at a site is not necessary for stable CRACLE results (Fig. 1.5). This is important for rapid collection of field data and successful application to fossil sites, both scenarios where taxa are likely to be missed. The primary distributional data for CRACLE are

also easily obtained, given the abundance of geographic collection data freely available via GBIF.

Potential caveats – A few potential caveats with regard to CRACLE should be explored in the future. The method assumes that the point-coordinate data for a species is a representative sample of the entire distribution geographically and, therefore, climatically. Thorough, nonbiased sampling across a species' distribution will provide the information necessary for an estimation of the realized niche (Hutchinson, 1957; Peterson et al., 2011). Nonrandom distribution of samples in geographic space may result in a data artifact that will influence CRACLE in a non–biologically meaningful way. It has been suggested that such sampling bias may be present at various levels in point-coordinate data based on specimen collections (Yesson et al., 2007; Feeley and Silman, 2011; Beck et al., 2014; Engemann et al., 2015), and for the present study we elected to exclude sites in Europe because of strong bias in occurrence across political boundaries, evident in GBIF. In future studies, correction of spatial bias may be shown to improve CRACLE performance and facilitate confident expansion into areas and taxa that exhibit strong spatial bias (Syfert et al., 2013).

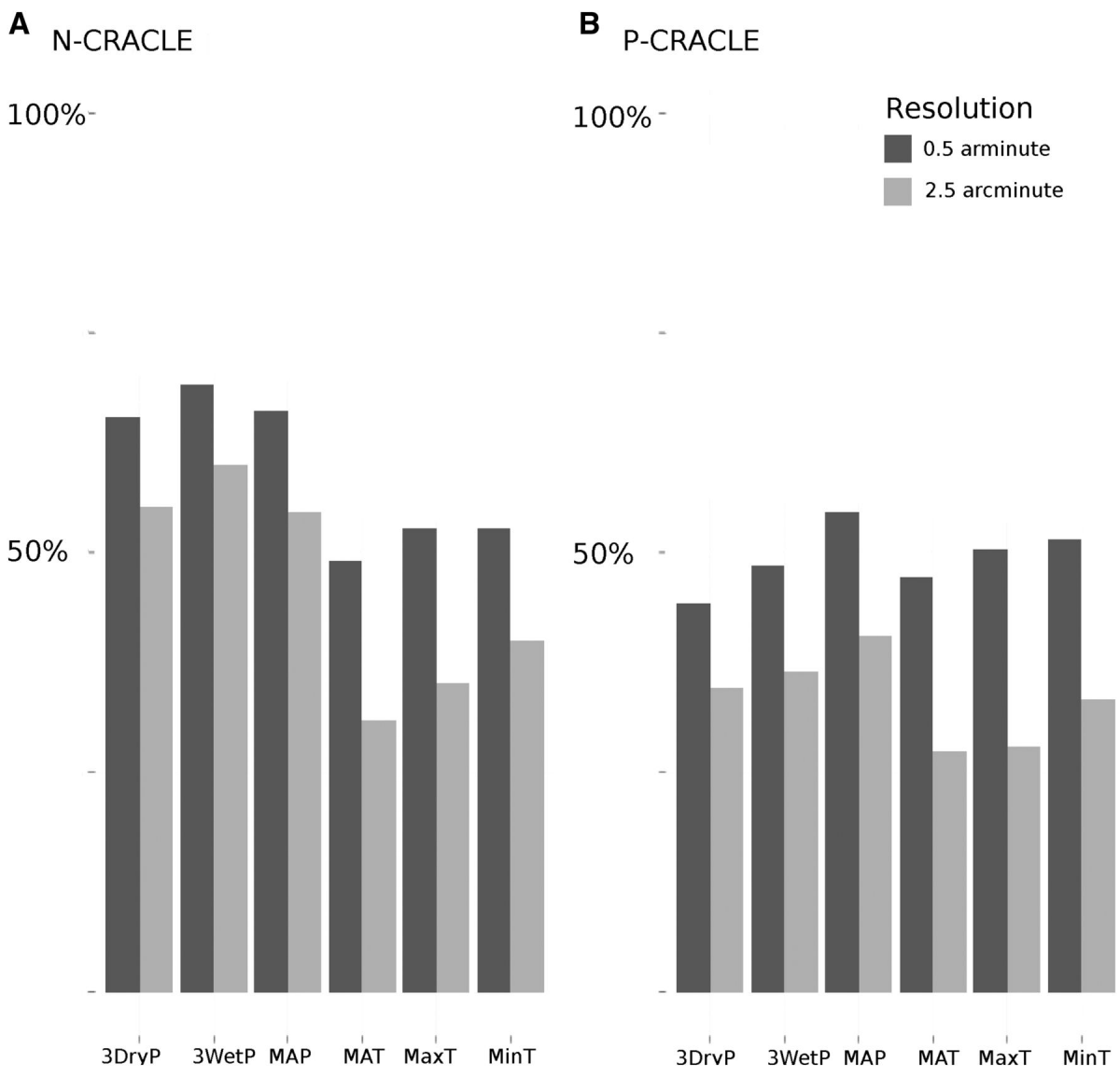

Figure 1.6. CRACLE overlap with WorldClim model output sampled at either 2.5 arcminutes, or local resampling of 0.5 arcminutes. The higher-resolution 0.5-arcminute WorldClim grid was resampled to summarize local climatic variation suggested by that model in a 2.5-arcminute square centered on each occurrence record. The range of climate values using the sample from the higher-resolution grid and the 2.5-arcminute WorldClim values were compared against CRACLE. The percentage of sites where CRACLE overlapped with each WorldClim estimate is reported for

N-CRACLE (A) and P-CRACLE (B) for each climate variable.

CRACLE fundamentally asserts that the climate with the maximum joint occurrence likelihood is where the species found together should be occurring. In reality, it is reasonable to expect that there are other, similar, climates where those species may co-occur that are slightly less “likely.” Sites that occur toward physical or climatic barriers to a vegetation type (e.g., the coastline or major shifts in precipitation regimes) may be situated systematically away from the center of distribution for a majority of the citizen species, skewing CRACLE results away from the actual climate and toward the more statistically likely climate. Identifying this hypothetical situation (an “edge” effect) both qualitatively and quantitatively should be explored to provide a methodological guide for the use of CRACLE in such situations.

Table 1.3. CLAMP, LMA, digital leaf physiognomy, and one coexistence approach (CA) model results statistics for temperature estimates.

	CLAMP (annual temperatures)			Royer et al., 2005 (mean annual temperature)			Punyasena, 2008 (annual temperature			
	Mean	Max	Min	Mean LMA	Mean Overall	Mean Fossil	Occurrence (mean, max, min)	Absolute abundance (mean, max, min)	Proportional abundance (mean, max, min)	Proportional basal area (mean, max, min)
Pearson's Correlation	0.91	0.87	0.76	0.89	0.94	0.93	--	--	--	--
Spearman's Rank-correlation	0.92	0.72	0.84	0.86	0.93	0.92	(0.48, 0.58, 0.654)	(0.561, 0.775, 0.594)	(0.47, 0.625, 0.457)	(0.281, 0.148, 0.22)
Mean error	1.8°C	2.1°C	3.2°C	2.0°C	1.9°C	1.9°C	--	--	--	--

Note: Spearman's Rank-Correlation values are reported from Punyasena (2008). All other values are based on simple analysis of published results (Royer et al., 2005; Jacques et al., 2011)

Table 1.4. CLAMP, and a coexistence approach (CA) model results statistics for precipitation estimates.

	CLAMP (precipitation)		Punyasena, 2008 (mean annual precipitation)			
	3 Wet	3 Dry	Occurrence	Absolute abundance	Proportional abundance	Proportional basal area
Pearson's Correlation	0.62	0.60	--	--	--	--
Spearman's Rank-correlation	0.59	0.57	0.39	0.27	0.12	-0.36
Mean error	157 mm	61 mm	--	--	--	--

Note: Spearman's Rank-Correlation values are reported from Punyasena (2008). All other values are based on simple analysis of published results (Royer et al., 2005; Jacques et al., 2011)

In many harsh environments, dormancy allows plants to escape seasonal climatic extremes and reduces sensitivity to the magnitude of seasonal drought or cold. For example, deciduous trees can completely avoid leaf damage due to freezing temperatures during winter and, thus, are less sensitive to the absolute minimum temperature (e.g., many of the same species have broad distributions from Georgia to Canada in eastern North America). A second example can be seen in the greater error and bias of CRACLE for estimation of the precipitation of the three driest months (Fig. 1.4E, F). One explanation is that plants exposed

to the most extreme drought are not “sensing” the severity of the dry season, owing to some form of dormancy. Any functional adaptation that releases a plant from climate as a continuous scale of selective force (e.g., escaping extreme cold or drought via dormancy) may limit the performance of CRACLE for predicting the more extreme climatic conditions. Alternatively, lifestyle strategies may limit the growing season for some species to very short spans of seasonal climate. For example, spring ephemerals complete their life cycle over the course of weeks or a few months when their environments are optimal for growth. In the eastern United States, this corresponds to a cool, wet season with plentiful light before the canopy has leafed out. These plants may be poor indicators of both winter and summer climate variables, owing to their focused spring growth. One way of objectively improving CRACLE performance in these areas may involve defining a set of phylogenetic, statistical, physiological, ecological, and/or morphological parameters for weighting species likelihood functions in relation to their sensitivity to particular climate variables and, consequently, their performance as climate indicator species. Investigation of this aspect of CRACLE will be important for future studies.

Table 1.5: CRACLE vs. WorldClim estimate Spearman's Rank-correlation

	Spearman's Rank-correlation	
	P-CRACLE	N-CRACLE
Mean annual temperature	0.96	0.96
Minimum annual temperature	0.91	0.91
Maximum annual temperature	0.98	0.98
Mean precipitation	0.89	0.90
Precipitation of the 3 wettest months	0.92	0.90
Precipitation of the 3 driest months	0.83	0.72

Conclusions

CRACLE yields a lower error rate in estimating climate than other ECV methods for which comparable validation analysis has been performed. This improvement in performance has been accomplished through the application of high-quality distributional data, high-resolution climate models, and empirical likelihood estimation based on explicit specimen collection data to the Coexistence Approach and related methods (see Mosbrugger and

Utescher, 1997; Kühl et al., 2002; Punyasena, 2008;Thompson et al., 2012). CRACLE arguably returns more consistent results not only compared to previous Coexistence Approach analyses (Mosbrugger and Utescher, 1997; Punyasena, 2008; Grimm and Denk, 2012; Thompson et al., 2012) but also to the commonly used physiognomic ECV methods (Wolfe, 1993, 1995; Wilf, 1997; Royer et al., 2005; Spicer et al., 2009; Peppe et al., 2011). CRACLE may also be highly sensitive to local climatic variation over very small spatial scales (Fig. 1.6) and therefore may be useful for climate model development and validation in areas with low weather station density. Despite these successes, it is important to point out that CRACLE does not yet compete with physiognomic ECV methods (Wolfe, 1993, 1995; Wilf, 1997; Royer et al., 2005; Spicer et al., 2009; Peppe et al., 2011) for the inference of paleoclimate. CRACLE has potential to be applied in that capacity for recent, Pleistocene and Holocene, paleoclimatic inference, where taxonomic placement of plant macrofossils or pollen can be made to modern species and realized climate niches assumed to have been relatively constant. Alternatively, more ancient climates may be investigated using CRACLE by incorporating phylogenetic hypotheses of fossil relationships and estimating fossil climatic niche characteristics based on related modern taxa (i.e., an enhanced Nearest Living

Relative method; Mosbrugger and Utescher, 1997).

CRACLE performs well across a wide variety of sites, from tropical to polar and desert to wet forest, supporting the hypothesis that plant distributions are independently correlated with climate and that the coexistence of species in a flora is highly predictive of the environment (here, the climate) that led to the formation of that community. The underlying hypothesis of vegetation assembly assumed by CRACLE is not new (see, especially, Gleason, 1926; Whittaker, 1956; Whittaker, 1967) and is difficult to demonstrate with certainty. The results presented here support the concept that plant communities are high-quality records of long-term climate. This suggests that CRACLE presents an opportunity for alternative assessment of interpolated climate models and for modeling paleoclimate.

Acknowledgements

We thank the following organizations for the data used in this study: www.gbif.org, www.salvias.net, www.worlclim.org, harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf116, and the other sources enumerated in Appendix S1.4. Approximately 3.75 million unique Global Biodiversity Information Facility (GBIF) records were used for this study, and for that we thank GBIF and all

of the data contributors whose enormous collective effort has made the current work possible. We thank H. Frye, A. Hill, T. Choo, and D. Allevato for assistance with data processing and quality control. We thank W. Crepet, A. Gandolfo-Nixon, I. Escapa, T. Choo, D. Allevato, B. Blonder, and several anonymous reviewers for their thoughtful comments on the manuscript.

CHAPTER 2

QUANTIFICATION OF CRACLE PERFORMANCE USING HIGHER TAXA: THE SMALLEST INCLUSIVE GROUP (SIG) AS A SPECIES SURROGATE.

Abstract

Application of CRACLE to fossil floras depends on finding modern analogs to the fossil taxa. The best analogs for physiology, and therefore climate tolerance, are likely to be closely related to the fossil taxa. Clades are often characterized by genetic, physiological, and morphological canalization that results in observed synapomorphies as well as similarities in ecological niches. Therefore, systematically closely related modern taxa are often used as a modern analog for fossil taxa. For convenience, higher taxa (e.g., at the genus level), have previously been used to this end. To test how CRACLE performs when using higher taxa to approximate species distributions modern vegetation survey data were analyzed using CRACLE with known species represented by their respective genera. Under these conditions CRACLE results do show slightly deteriorated performance, but primarily only when more than half of the species at a locality are represented by

genera. For example, the estimation of Mean Annual Temperature under the Gaussian CRACLE method has mean anomaly rate of 1.44°C when only species are used. This increases to 2.03°C when all taxa are represented as genera. These results provide a baseline estimate for how CRACLE can perform when using species surrogates for extinct fossil taxa.

Introduction

Estimation of paleoclimate with Climate Reconstruction using Coexistence Likelihood Estimation (CRACLE) is an obvious next step in development and application of the method. To utilize CRACLE with fossil data, either fossils must represent living species, or a surrogate taxon selected from modern taxa must be used. If species at a fossil locality are extinct, it is not possible to utilize a single modern plant species to characterize the estimated niche for that fossil. The most common approach to dealing with the issue of extinct species in paleoclimate reconstruction has been the nearest living relative (NLR) method as applied routinely in the Coexistence Approach (CA – Mosbrugger and Utescher, 1997; Utescher et al., 2014). The NLR is defined as “systematically closely related” (Mosbrugger and Utescher, 1997) to the fossil taxon. However, details of surrogate (NLR) selection have not been

adequately described in studies using the method. For example, the Paleoflora database provides users with proposed NLRs for many Tertiary fossil taxa without a clear framework for how those NLRs are identified (Utescher and Mosbrugger, 2015). For convenience higher taxa are often used in CA-like (e.g., Boyle et al., 2008; Thompson et al., 2012) NLR climate reconstruction methods. With this approach, if a fossil clearly fits within a modern genus or family based on cladistic analyses, then the entire group is used as a surrogate to characterize the climatic or ecological preferences of the fossil taxon. It has recently been suggested that the NLR should be phylogenetically determined, or at least care should be taken if using higher taxonomic groups. Furthermore, quantitative validation of CA-like methods (including CRACLE) should be studied further using higher taxonomic groups in modern floras in order to provide a robust validation of the approach to paleoclimate reconstruction (Grimm and Potts, 2015).

For our purposes, we propose the concept of “smallest inclusive group” (SIG) as a surrogate to characterize an extinct species. This approach allows, but does not require, using an entire modern genus to characterize an extinct species. The SIG, is defined here as the smallest monophyletic group (or taxonomic group, if a phylogeny is not available) that includes the species in

question. SIG is a distinct, but similar method to NLR. Using the climate tolerances of the NLR/SIG (be it a genus, species, clade, or analog) requires an assumption of sufficient physiological, and therefore climatic, uniformitarianism to allow the surrogate group to adequately represent the extinct species (Tiffney and Manchester, 2001; Tiffney, 2008; Grimm and Potts, 2015). Whether or not, and the degree to which taxa maintain sufficiently stable realized climate niches through time have been the subjects of much debate (Araujo and Guisan, 2006; Sexton et al., 2009; Wiens et al., 2010; Araujo et al., 2013). For some species, it is expected to be true that the realized climate niche may change rapidly through time. However, the realized (and fundamental) niches of two sister taxa are likely to be more similar to one another than to a randomly selected species (e.g., Covreur et al., 2011), implying phylogenetic niche stasis (to some degree) or at least stability. Niche stability and tracking through time has been observed in the recent (Late Quaternary) fossil record (Prentice, 1991), but other analyses have observed measurable lags in niche tracking (Orodonez, 2013). The assumption of sufficiently stable realized climate niches is often made in other related methods of paleoclimate analysis (e.g. Mosbrugger and Utescher, 1997; Sinka and Atkinson,1999; Kuhl et al. 2002; Punyasena, 2008; Thompson et al. 2008; Thompson et al.,

2012; and all applications of these methods).

Assuming stable realized niches, there remains the question of how to best define “systematically close” in an NLR analysis (Grimm and Potts, 2015). The preferred approach to this would be to use phylogenetic analysis to place the fossil taxon within an extant clade and then to use the sister taxon or smallest inclusive monophyletic clade to approximate the fossil taxon's niche parameters. However, data required to accurately place a fossil in a phylogenetic context may not always be accessible, and/or results of such analyses may be ambiguous. An alternative to specific phylogenetic placement may be identifying the smallest taxonomic group (i.e., genus) that includes the fossil based on the presence of that taxon's diagnostic characters, which should approximate a phylogenetic placement assuming that modern taxonomy is based on phylogenetic analyses.

Grimm and Potts (2015) provide a hypothetical example wherein two co-occurring (overlapping) fossil taxa will have narrower ranges than their respective genera and therefore the use of genera to define NLRs and the coexistence interval will result in a broader (less precise) estimate of the climate envelope. Building CRACLE likelihood profiles with generic distributions will result in broader overall niche dimensions, but the shape (preference) of the

niche will be influenced most by the most widespread or most collected specie(s) in that group. In the likelihood framework of CRACLE characterizing the genus this way makes sense; if the placement of the fossil species within the group is unknown, then the least biased assumption is that it is probably like the more common specie(s) in the group. Therefore, given the likelihood distributions of two overlapping genera the likelihood functions of two overlapping species included in those groups have the strong potential to preserve a reasonably accurate maximum likelihood value. This potential becomes more probable as more overlapping genera are added, thereby excluding climate space unoccupied by some taxa (the same is true for species with distinct populations or broad distributions).

It has been argued that the use of higher taxa may be a misleading way to identify NLR climate tolerance primarily because it is possible for two species to not overlap in their climate tolerances even though their respective genera do overlap (Grimm and Potts, 2015). In fossil deposits, two non-contemporaneous fossil taxa may appear to be in the same fossil community, therefore a taxonomic paleoclimate reconstruction method (i.e., CRACLE) should be robust to outlier taxa (non-overlapping taxa) and this will become more difficult when using higher taxonomic groups. In

CRACLE there is no probability of occurrence of zero (log-likelihood = -Inf), every possible value of a parameter has a finite, non-zero probability. Normally, very low likelihoods will result in those values of climate being excluded as possible choices for the model. In the case of two non-overlapping (by coexistence intervals) taxa this principle would place the maximum-likelihood value approximately halfway between the two non-overlapping taxa assuming symmetric distributions. When more taxa are available in the model, the essentially non-overlapping taxon (the outlier) will have a negligible effect on the maximum likelihood value.

CRACLE provides a framework that may be robust in the face of these potential issues associated with fossil taxa; however, when broader taxonomic groups such as genera are used as extinct species surrogates the inferred climate tolerance will be inherently broader. The question remains: How does the use of higher taxonomic groups change the accuracy and precision of CRACLE? The goal of this experiment is to quantify the expected effects on CRACLE performance using modern vegetation surveys to generate simulated data sets where only species are sampled vs. progressively more and more taxonomic uncertainty represented by coding defined percentages of species as their respective genera (SIGs) until all taxa are represented as genera. A variety of

performance statistics will quantify model performance as taxonomic uncertainty increases and also examine correlations with external factors (e.g., geographic regions, diversity of vegetation).

Methods

The 353 modern data sets, including those assembled in Harbert and Nixon (2015), were used to simulate data sets with varying percentages of surrogate species based on generic level smallest inclusive groups. As a conservative approach, and to avoid the necessity to generate phylogenies for every species, species were incrementally replaced by genus-level niche profiles.

CRACLE – Estimation of climate based on species coexistence and modern species distributions via the Climate Reconstruction Analysis Using Coexistence Likelihood Estimation (CRACLE) protocol (Harbert and Nixon, 2015). This method generates parametric (normal Gaussian) and non-parametric (Gaussian Kernel Density Estimation) probability functions for the occurrence of a species along a dimension of climate (e.g., average annual temperature). The joint likelihood function for all co-occurring species is then calculated as the product (or sum-log-likelihood) of these species functions. The maximum of the joint

likelihood curve is taken to be the most probable climate value given the association of species and their individual association with climate.

For this study, CRACLE was implemented (CRACLE Script v2.0 can be found in Appendix S2.2) in the previously described manner with only two changes. To optimize the Kernel Density Estimation procedure Silverman's Rule was applied to select the near optimal bandwidth (Silverman, 1986) rather than using predefined bandwidth for each variable (Harbert and Nixon, 2015). Second, to better characterize the precision of the CRACLE estimate 95% confidence intervals were calculated for each joint likelihood distribution (the middle 95% of the probability density function). Mean and median anomalies are based on the maximum likelihood value but these intervals were analyzed separately to quantify model precision effects.

Using higher taxa –To quantify the effects of using higher taxa to approximate the smallest inclusive group as a surrogate species for extinct fossil taxa each of 353 modern test sites was analyzed seven ways: first, with all known species, then with 20%, 35%, 50%, 65%, 90%, and 100% of those species coded as their respective genera. Each site was analyzed only once for each condition. The

choice of which taxa to translate to genus were made at random using a random number generator to select a predefined proportion of the taxa for translation.

The GBIF data were filtered so that only one record was included from any one 2.5 arcminute WorldClim grid cell, to reduce spatial sampling bias that does not reflect the natural distribution and density of a species (i.e., multiple collections of a species in the same locality). In the case of generic distributions the same rule was applied the pooled distribution records for all species in the genus.

Modern climate model – Climate data are taken from the downscaled 2.5-arcminute resolution (~0.041667 degrees) WorldClim model grid (Hijmans et al., 2005). WorldClim is a high-resolution continuous grid of interpolated climate data for the world's land areas derived from >40 000 weather stations around the world.

Climate Variables – One of the major advantages of the CRACLE method for inferring paleoclimate is that the general protocol is flexible and can be applied to a wide variety of climatic parameters. Descriptions of the variables analyzed for this study

can be found in Table 2.1. For the first time here, to better capture nuances of climate relating to drought, the potential evapotranspiration (PET) and water balance (precipitation – PET) was calculated using monthly values for temperature and precipitation from the WorldClim model (Hijmans et al., 2005). The Thornwaite Equation (Thornwaite, 1948) was chosen as a suitable model of potential evapotranspiration that relied on data available at the scale of WorldClim including: 1) monthly average temperatures in degrees Celsius, 2) day length in hours (calculated from the latitude and month), 3) the number of days in each month.

Table 2.1: Climate variable definitions and units

Variable Name	Description	Units
MAT*	Mean Annual Temperature	°C
MaximumT*	Average maximum temperature of the warmest month	°C
MinimumT*	Average minimum temperature of the coldest month	°C
tempbalance	Sum of warm degree days (days > 0°C)*mean monthly temperature and cold degree days (days <0°C)*mean monthly temperature. -Such that 0 correspond closely to MAT = 0°C, but other values represent the relative proportion of warm season vs. cold season.	°C
diurnal	Mean diurnal temperature change	°C
GSL	Growing season length -Months when DRLEN and WINTERLEN (both defined below) conditions are both not met.	months
MAP*	Mean Annual Precipitation	mm
GSPREC	Precipitation of the GSL associated growing season	mm
wbalann	Annual water balance	mm

	-Mean Annual Potential Evapotranspiration - Mean Annual Prec. Note: >0 suggests more water comes in than theoretically evaporates	
maxwbal	Maximum monthly water balance	mm
minwbal	Minimum monthly water balance	mm
wet_mo	Wet months -Months where water balance > 0	months
wet_mo_mean	Average Wet month water balance	mm
wet_sum	Sum water balance of the wet_mo defined wet season	mm
dry_mo	Dry months -Months where water balance < 0	months
dry_mo_mean	Average dry month water balance	mm
dry_sum	Sum water balance of the dry_mo defined dry season	mm
X3DryP*	Precipitation of the 3 driest months	mm
X3WetP*	Precipitation of the 3 wettest months	mm
DRLEN#	Drought length -Months where ($temp_i$ - ($prec_i$/2)) >= 0	months
DRSEV#	Drought severity - DRSEV(drought months 1:i) = sum($temp_i$ - ($prec_i$/2))	unit-less
WINTERLEN#	Winter length -Months when the mean monthly temperature is <5°C.	months

*Unmodified from the WorldClim Bioclim variables (www.worldclim.org/bioclim; Hijmans et al., 2005)

After: Walter, 1973 .

Results

When the simulations were run with 22 climate variables (Table 2.1) there was usually no significant effect on the accuracy of CRACLE when fewer than 50% of the taxa were coded as genera (Tables S2.2 and S2.3). For Mean Annual Temperature (MAT) and Mean Annual Precipitation (MAP) these trends are shown more closely below as representative of the robustness of CRACLE when estimating temperature and precipitation trends respectively. In all cases, the absolute anomalies are skewed right (towards zero) so the median may be a better representation of the central value and the mean anomaly may be sensitive to outliers (poor model results). To examine how the use of higher taxa affects performance of CRACLE as measured by these parameters 95% confidence intervals of the mean and median were generated from 9999 bootstrap replicates sampling 60% of the 353 results (Fig 2.1,2.2,2.3).

Both MAT and MAP show general deterioration of model performance as more taxa are replaced by their generic distributions as surrogates for the simulation of unknown species distributions. For mean anomalies of MAT prediction the Gaussian CRACLE results suggest that this difference is not significant until 65% of the taxa are represented as genera, where the KDE CRACLE (identical to N-CRACLE in Chapter 1) results suggest that the mean anomaly difference is significantly increased after only 35% of the taxa are sampled as genera (Fig. 2.1). MAP on the other hand has robust estimates using the KDE CRACLE method until 90% of the taxa were substituted with genera, vs. again 65%

using the Gaussian CRACLE (identical to the P-CRACLE method in Chapter 1) method (Fig. 2.2).These results are broadly supported by multiple comparison hypothesis testing using the Games-Howell method (Games and Howell, 1976; Day and Quinn, 1989) (Table 2.2). Model precision, as measured by average estimated range, is affected by the use of higher taxa but the deviation is not significant until at least 90% of the taxa are coded as genera (Table 2.3).

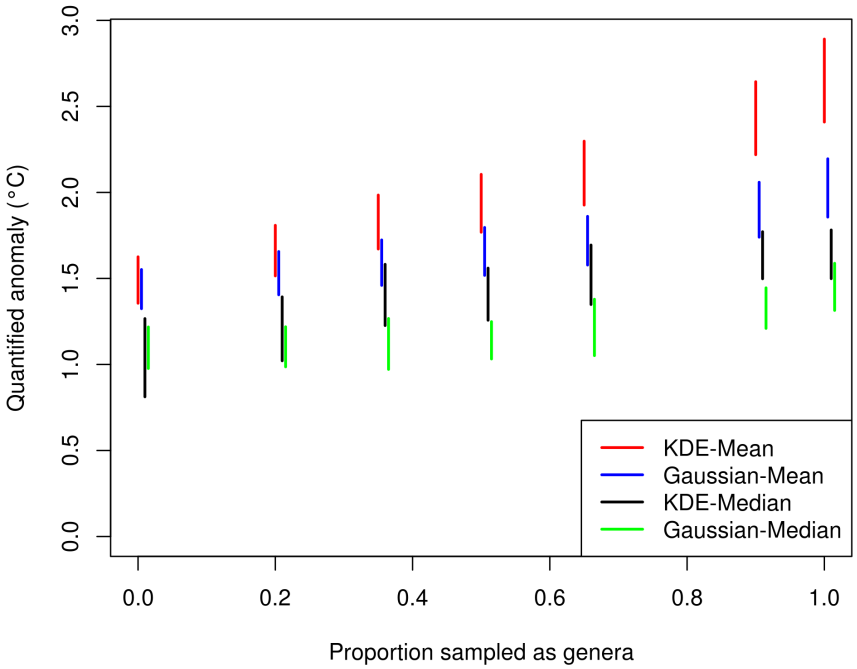

Figure 2.1. Mean Annual Temperature (C) mean and median anomaly confidence intervals.

Median anomalies of MAT are less affected than mean anomalies by the taxon sampling strategy (vs. mean anomalies). MAT median anomalies under the Gaussian CRACLE method show no significant increase until greater than 90% of the taxa have been sampled as genera (Fig. 2.1) using the bootstrap method. Precipitation (MAP) shows a different result;

median MAP anomalies under the Gaussian CRACLE method show significant increases after 50% of the taxa were sampled as genera (vs. 90% for the mean anomaly under the same conditions) (Fig. 2.2).

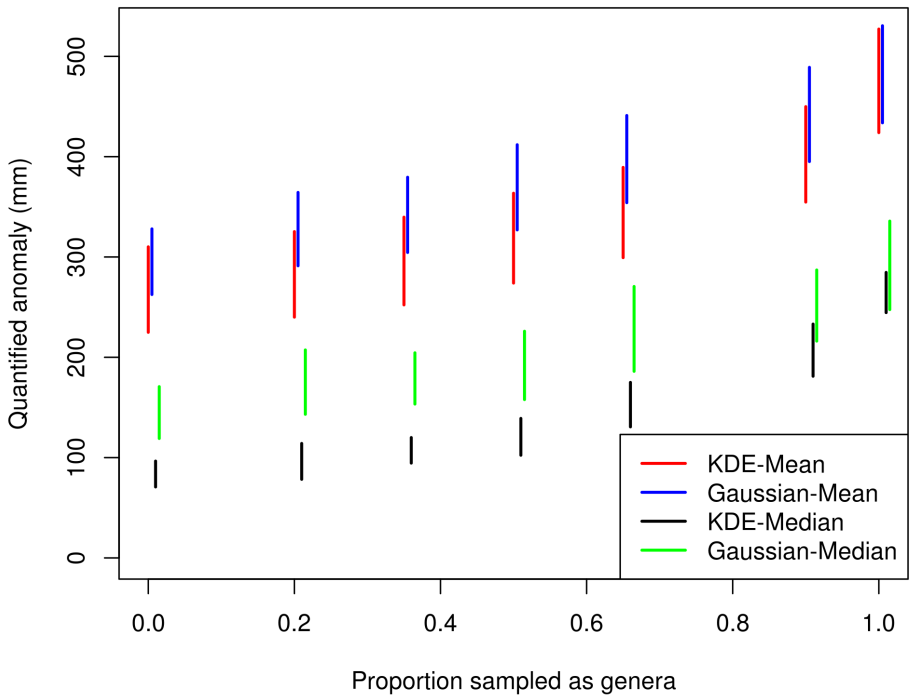

Figure 2.2. Mean Annual Precipitation (mm) mean and median anomaly bootstrap confidence intervals.

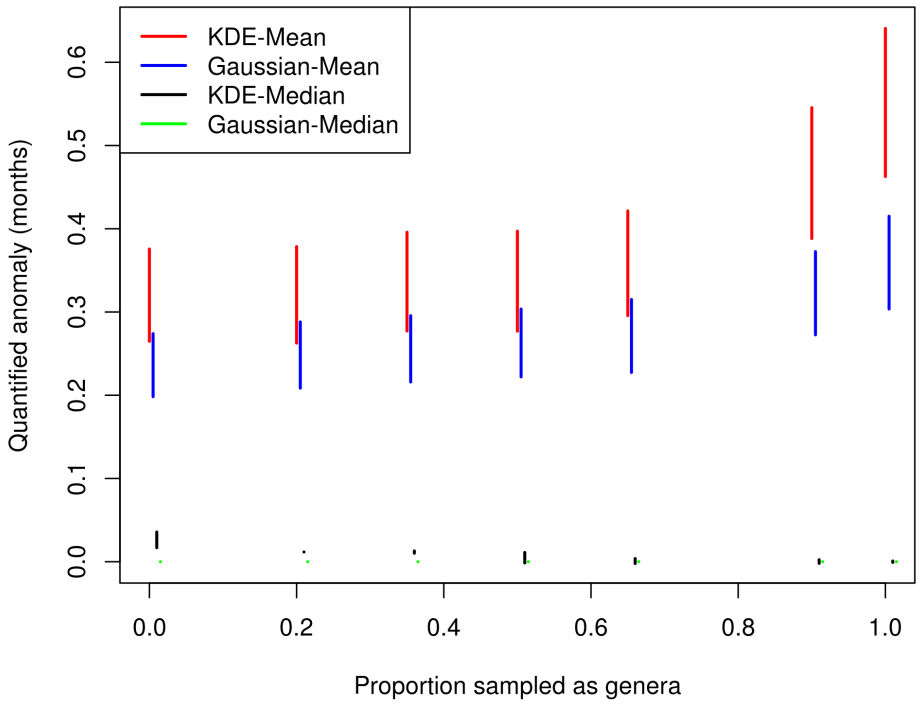

Figure 2.3. Winter Length (months) mean and median anomaly confidence intervals.

Estimation of winter length using either method of CRACLE was unique in that there was no significant increase in the median anomaly (which was 0 in all cases) and a very small increase in mean anomalies was observed up to only about 0.5 month average anomaly when using all genera (Fig. 2.3). The estimation of winter length is the highest performing seasonal parameter analyzed with CRACLE (Appendix Table S2.2 and S2.3).

Table 2.2: Multiple Comparison testing of mean anomalies using the Games-Howell method

% Genera	MAT vs. all species set		MAP vs. all species set	
	KDE	Gaussian	KDE	Gaussian
20%	0.81	0.98	1.00	0.93
35%	0.12	0.80	0.99	0.73
50%	**0.02**	0.44	0.84	0.29
65%	**0.00**	0.17	0.42	**0.04**
90%	**0.00**	**0.00**	**0.01**	**0.00**
100%	**0.00**	**0.00**	**0.00**	**0.00**

Table 2.3: Multiple Comparison testing of estimate ranges using the Games-Howell method

% Genera	MAT vs. all species set		MAP vs. all species set	
	KDE	Gaussian	KDE	Gaussian
20%	1.00	1.00	1.00	1.00
35%	1.00	1.00	1.00	1.00
50%	0.92	0.85	0.96	1.00
65%	0.88	0.35	0.11	0.87
90%	0.27	**0.00**	**0.01**	**0.01**
100%	**0.02**	**0.00**	**0.01**	**0.00**

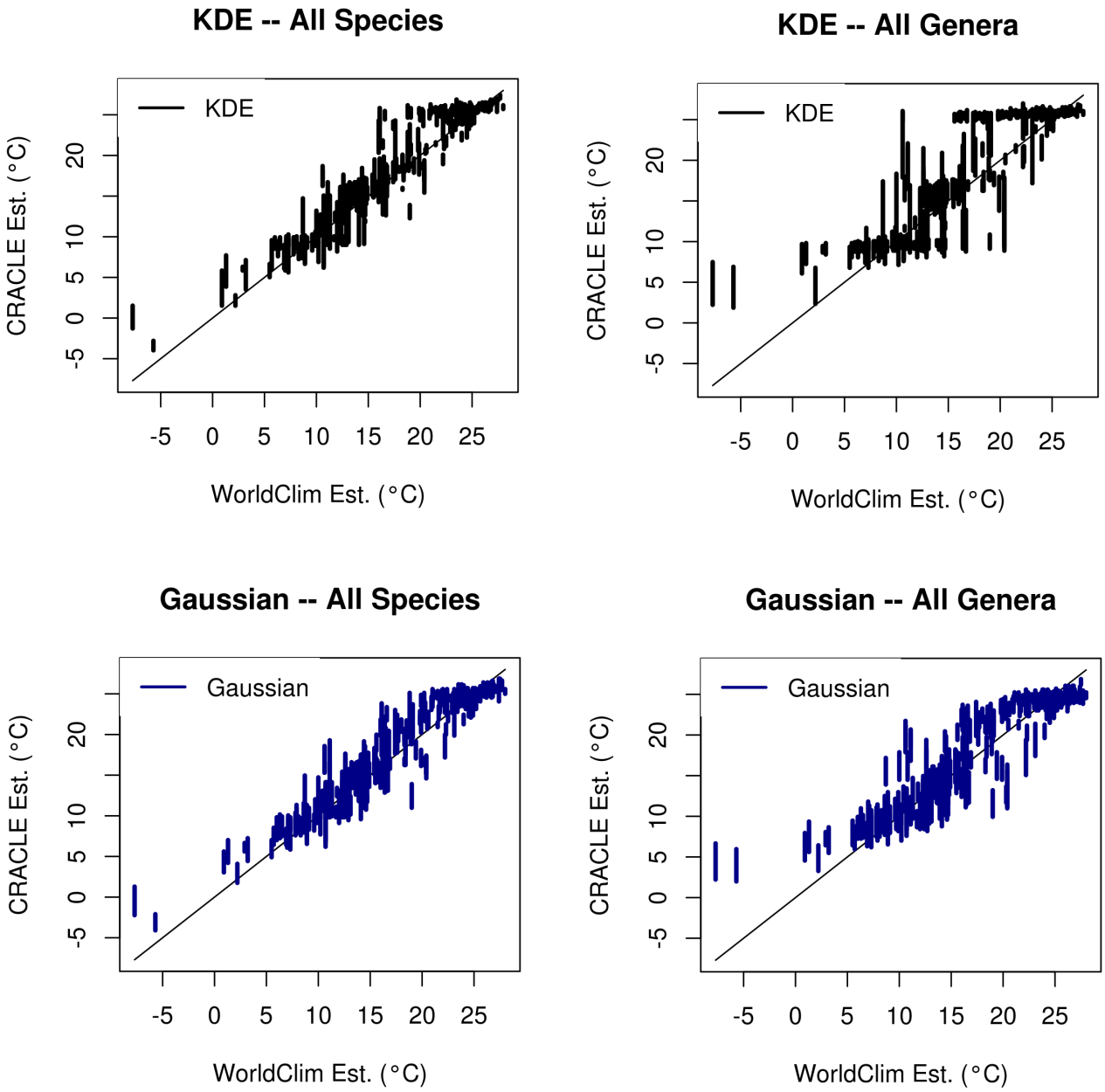

Figure 2.4. CRACLE model output for the estimation of Mean Annual Temperature (MAT) in C for 353 modern sites using taxa coded as species (A,C) or Genera (B,D).

Qualitative examination of the results data for each sample partition show that the KDE and Gaussian models are affected very

differently by generic surrogates (Fig. 2.4). In the KDE CRACLE models, the estimates lose accuracy such that the estimates converge on (generally) 3 levels (in temperature corresponding to “tropical”, “warm-temperate”, and “cold-temperate”). The Gaussian CRACLE method appears to be more robust to this effect and is more likely to have a lower anomaly rate when higher taxa are used (Fig. 2.1,2.2, and 2.3; Appendix Tables S2.2, S2.3. S2.4 and S2.5) and retain higher predictive power as measured by r2 values (Appendix Tables S2.6 and S2.7); though for some variables the KDE CRACLE method does retain higher performance (e.g., Mean Annual Precipitation – MAP –Fig. 2.2).

Discussions

Review of results – Measurable differences, i.e., increased error rates, are observed on a site-by-site basis as higher percentages of species are sampled using the changed to genus surrogates. However, these decreases in performance are not significant for most climate variables when fewer than 60% of the species are replaced by surrogate genera.

Taken together, these results address recently published concerns and critiques of taxonomic climate reconstruction methods in general, of which CRACLE is one subset (Thompson et al., 2012; Grimm and Potts, 2015). Concern has been expressed

over whether or not the use of higher taxa to approximate fossil climate tolerances will lead to unavoidable bias and errors (Grimm and Potts, 2015). CRACLE is not, on average, sensitive to replacement of less than 50% of the sampled known species with generic surrogates to approximate a smallest inclusive group (SIG) approach. When all species niches are approximated by their respective surrogate generic distributions the performance of CRACLE is significantly decreased, though the use of Gaussian probability functions (vs. the Kernel Density Estimation method) in CRACLE is more robust to this effect. The use of higher taxa (i.e., the genus) distribution to approximate the climate tolerance of a fossil species has been criticized for relying on niche uniformitarianism (Grimm and Potts, 2015) even though generic niches have been observed to be more stable through time (Huntley et al., 1989; Ackerly, 2003; Hadly et al., 2009; Wake et al., 2009) suggesting some sort of niche canalization at that level owing to physiological constraints. In this modern data set the systematic closeness of genera to the known species appears to be sufficient within the CRACLE framework to produce relatively robust estimations of climate when sufficient sample sizes are used.

The use of generic distributions to characterize known species in the CRACLE model results in an overall increased error rate for all variables tested (See Appendix Table S2.2:S2.5) with the

exception of the median anomaly for the estimation of winter length (WINTERLEN) which is not shown to increase significantly by either CRACLE method (Fig. 2.3). Although, mean anomalies for Winter length do increase slightly.

Model performance expectations when using higher taxa – The increased error rates observed with increasing use of generic surrogates are expected because generic distributions in climate space will almost always be broader than a species distribution except in cases of monotypic genera. The effect of broadening all taxon curves (by using generic surrogates for known species) is that the discriminatory power of the maximum likelihood function should be reduced because the likelihood curve is flattened: the maximally likely value is reduced and the likelihoods of nearby values relatively increased. Single species curves will generally be narrower and taller, as opposed to generic curves that will be lower and broader. This translates into lower model performance, particularly at the non-zero climatic margins (e.g., low temperatures or high precipitation). Combining lower, flatter, curves will have less discriminatory power and therefore more ambiguous results. Zero values of climate (e.g., very low precipitation or no drought/winter seasons) are less affected because there is no climate space beyond that boundary, so the

likelihood profiles cannot be pulled any farther away in that direction by broadening the species distributions with generic surrogates.

Another informative approach is to consider the extreme cases. If one were to identify a fossil only as “plant”, without a family, order, or other taxonomic designation, then the distribution of all plants could be considered as the SIG. In this case, considering for simplicity the Gaussian CRACLE model, the maximum likelihood value would be exactly at the global mean value for every climate variable (but this would likely be a very broad, not “sharp”, peak). With no precision on the taxonomy of a fossil, CRACLE is left with the maximum inaccuracy. The question is: How much better is CRACLE performing than the “mean-only” model using species and genera respectively? The r^2 value calculates this metric exactly. An r^2 of 0 corresponds to a model that performs no better than simply taking the mean of the dependent variables (in this case: the WorldClim estimates for the sample sites). For CRACLE, if the empirical model output for a sample of sites has an error rate proportional to the difference between each WorldClim value and the mean WorldClim value for all of those sites, then the r^2 would be 0. Alternatively, if every CRACLE result exactly matched the expected WorldClim estimate then the r^2 would be 1.

For some variables under the Gaussian CRACLE model (e.g., MAT, MinimumT, WINTERLEN) the r^2 values when all genera are used still suggest highly appropriate models (e.g., r^2 >0.8; Supplementary Tables S2.6 and S2.7). This is not true for any of the KDE CRACLE models, though some have r^2 values that indicate respectable models (e.g., r^2 >0.5 for the same variables indicated above). However, in both methods variables MaximumT, the water balance related variables, and X3DryP (precipitation of the driest 3 consecutive months) are examples of variables where r^2 suggests that the model is approaching uselessness (r^2 significantly < 0.5). In no cases are there 0, or negative r^2 values suggesting that there is always some signal, no matter how obscured, produced by the CRACLE model. These are anecdotal quantifications of how useful the model might be. However, r^2 , for the reasons outlined above, is a valuable metric for analyzing the usefulness of these models. In the results of both CRACLE models applied to all variables there are portions of the climate space that are well characterized. However, the use of higher taxa does impinge on model performance for many variables when the classic “edge effect” is in play (Fig. 2.4). The “edge effect” is the decrease in accuracy of CRACLE when a species association occurs at or near the edge of the correlated species distributions in geographic or climatic space. In such cases the maximally likely value of climate will be shifted

away from the actual locality, towards the center of the distribution of that association (Harbert and Nixon, 2015).

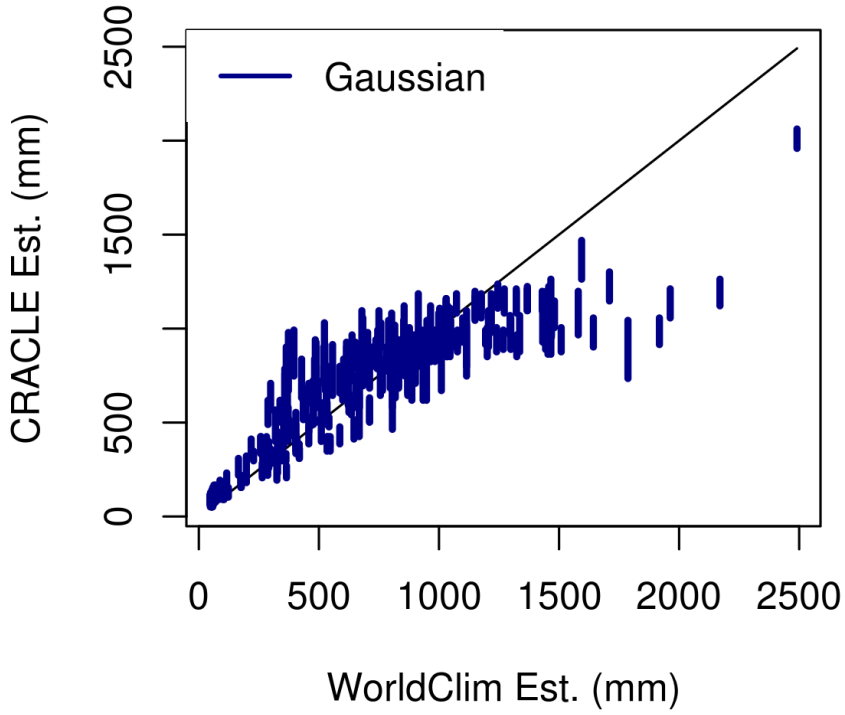

Figure 2.5. Estimation of the mean precipitation of the 3 wettest months by CRACLE exhibits a non-linear association with the WorldClim estimates.

Non-normality of CRACLE vs. WorldClim regression residuals

– Figure 2.5 shows an example of the individual anomalies of the CRACLE maximally likely intervals vs. the WorldClim estimated values for X3WetP (Precipitation of the three wettest months). Interpretation of this particular model's performance is not well characterized simply by looking at the magnitude of the anomalies (119.35 mm mean; 61.18 mm median) or the r^2 (0.74). This is because the anomalies are not evenly distributed with respect to the expected (WorldClim) values when those values are high. As the WorldClim expected values increase above 1000 mm the anomalies also increase. These results suggest that the CRACLE models are still sensitive to the variation in this climatic parameter but the accuracy decreases proportionally to the WorldClim estimated value. This suggests that one avenue of model improvement would be to build polynomial (or otherwise best fitting) regressions to a test set of CRACLE model output to quantify the biased error rates relative to the WorldClim estimation. These functions could then be used on independent CRACLE models to adjust model output to account for the magnitude and direction of these known error biases. These correction functions would help to improve CRACLE estimates on the climatic extremes where the magnitude and direction of error is significant but potentially predictable (i.e., in extreme “edge effect” cases).

Moving Forward –Given these results how should one proceed with CRACLE analyses in a paleobotanical context, where species may be extinct or inadequately identified? Using CRACLE “as-is” still provides a suitable model for using paleo-floras to estimate climate when fossil species can be identified as belonging to modern species or genera. Care should be taken with the situations (variables and models) where increased model error could result from the use of higher taxa as surrogates. For example, in the recent past (Late Pleistocene and Holocene) it is often possible to associate fossil plant taxa with modern taxa as there are no, or minimal, morphological differences between the fossil and modern specimens (Betancourt et al., 1990). Physiological uniforminatarianism may not be absolute due to the disconnect between realized and fundamental niches (e.g., Araujo et al., 2013) but should be expected as the majority rule during this period (Prentice, 1991). On average, it is expected that if niches are not uniform, then for a given association of species individual taxa should be shifting niches independently from one another and this effect should “wash out” of the CRACLE results.

The use of higher taxa (families or genera) as SIG surrogates with CRACLE will require more future testing to provide “best practices” guidelines. Some steps that could improve CRACLE

performance when using higher taxa as SIG surrogates are the use of regional geographic restriction of surrogate distributions (if applicable to a fossil flora in consideration of biogeography) or morphological restriction (e.g., the “convexly lobe-leafed-oaks” vs. the entire genus *Quercus* could provide a more refined climate tolerance profile). Such restrictions, if reasonably justified, could produce narrower, more decisive climate curves that would result in more precise estimates. Additionally, other phylogenetically based approaches could be developed to estimate climate niches for fossil taxa based on optimization of climate variables in the context of tree structure. Such approaches could be used to generate more specific hypotheses of climate tolerance (character) evolution to better characterize the fossil taxon likelihood profiles.

Conclusions

The development and refinement of CRACLE and supporting methodology has the potential to yield a method that will be broadly applicable to fossil floras through the Tertiary and, with some caveats, to older fossil assemblages. The experiments presented here establish a preliminary foray into the use of higher taxa as species surrogates in the CRACLE model and suggest that this approach is generally robust to inclusion of up to 50% surrogate taxa, and in some cases higher. The use of CRACLE with a mixed

taxon model consisting of either genera or species, depending on what is known about a fossil, should be able to produce high quality estimates of climate when fossil taxa are very closely related to modern groups. CRACLE, therefore, is a robust and feasible approach to characterizing paleoclimate of underrepresented regions in the Late Quaternary climate record of the mid-latitude terrestrial environments that are often well-represented in the macrofossil and pollen records through this time.

CHAPTER 3

A NOVEL 50,000 YEAR ESTIMATE OF CLIMATE CHANGE INFERRED FROM PACKRAT (*NEOTOMA* SPP.) MIDDEN PLANT FOSSIL RECORDS IN WESTERN NORTH AMERICA.

Abstract

CRACLE is recently developed approach to the estimation of climate from local plant taxonomic diversity data, and has been validated as highly accurate using modern plant distributions and published climate models based on modern weather records. This study reports on the application of CRACLE using plant macrofossil identifications from packrat (*Neotoma* spp.) middens of western North America. The results are a ~50,000 year quantitative estimation of late Pleistocene and Holocene paleoclimate of western North America. The CRACLE estimates include a wide array of variables including temperature, precipitation, available moisture, and seasonal patterns and are consistent with well-understood climate forcing factors such as variations in summer insolation and CO_2 concentration. This climate reconstruction also corroborates many well-understood climatic patterns throughout this period including the terminal Pleistocene glaciation, Younger Dryas, Holocene Thermal Optimum, and the late Holocene cooling trend.

Introduction

Strong evidence of significant glaciation during the Quaternary has been observed for >150 years (Agassiz, 1840; Carozzi, 1967), culminating in the last major glacial maximum at about 21kya (thousand years ago). These periodic glaciation – deglaciation cycles appear to be driven by predictable and periodic variation in the earth's orbit relative to the sun (Milankovitch, 1930; Berger et al. 1978) and the effects that this variation in sun angle and intensity have on the earth's climate system (review: Imbrie et al., 1992). The global glacial record is well preserved and trapped within the ice are atmospheric gases which can provide stable isotope proxies for temperature, in the Greenland and Antarctic ice-sheets (Cuffey and Clow, 1997; Alley, 2000; Alley, 2004; Bereiter, et al. 2015). Lower latitude quantitative climate records are less continuous and primarily limited to ocean and lake sediment (e.g., Barron et al., 2003, MacDonald et al., 2008) and speleothem/cave groundwater stable isotope analyses (e.g., Marcott et al., 2013).

The modern climate of the study area – the American Southwest – is characterized by the North American Monsoon (Adams and Comrie, 1997; Wright et al., 2001). Global atmospheric circulation is affected by the heating of the North American continent, driving the warm season development of a low pressure

system across the Southwestern United States, resulting in increased summer precipitation across the region following a hot and dry spring and early summer (Wright et al., 2001). Heating of the atmosphere across the continent is driven by global temperatures and circulation (Wright et al., 2001) as well as by summer insolation which varies predictably with the earth's orbital pattern and is correlated with global temperature (Milankovitch, 1930; Berger et al. 1978; Imbrie et al., 1992).

During the last glacial period, cooler temperatures and the presence of an ice sheet to the north, resulted in the shift of the summer monsoon cycle to the south, along with the northern jet stream drifting south (Thompson et al. 1993; Oster et al., 2015). Due to these shifts, that total regional precipitation increased, presumably to a greater degree in the winter and spring, probably driving the development of large pluvial lakes in the Great Basin during the terminal Pleistocene glacial period (Brackenridge, 1978; Ibarra et al. 2014) and resulting in shifts in vegetation zones to the south and to lower elevations (Betancourt et al., 1990; Cole, 2009).

Evidence from the Greenland Ice Sheet Project 2 ice core (GISP2 – Cuffey and Clow, 1997; Alley, 2000; Alley, 2004) shows an anomalous North Atlantic regional cooling event ca. 12.9 and 11.7ka, known as the Younger Dryas, that drove vegetational shifts (e.g., the return of the cold-tolerant *Dryas octopetala* to the pollen

record of central Europe from interglacial refugia to the north), sea surface temperature cooling, and Northern Hemisphere glacial advances (Carlson, 2013). The effects of the Younger Dryas may have contributed to a temporarily cooler and wetter climate across the American Southwest (MacDonald et al., 2008). Evidence exists for sea surface depression of up to 4°C off of Northern California (Barron et al., 2003). Altitudinal shifts in *Agave utahensis* populations in the Grand Canyon (Cole and Arundel, 2005) have also been attributed to the Younger Dryas, suggesting that vegetation composition was likely affected throughout this region. The shift in *A. utahensis* populations to lower altitudes has been hypothesized to correspond to the Younger Dryas being ~8°C cooler than the modern, followed by rapid warming towards the Holocene Thermal Optimum (Cole and Arundel, 2005). Stable isotope analysis of bat guano from the Grand Canyon provides evidence for a warming and drying trend initiated about 15kya but with pauses or brief reversals corresponding to the Younger Dryas and an un-named event at ca. 8.2ka (Wurster et al., 2008).

The onset of the Holocene is characterized as a period of rapid warming as deglacial forcing drove rapid increases in global temperature. Northern Hemisphere summer insolation reached the maximum of a high-amplitude cycle around 10ka (Berger, 1978; Imbrie et al., 1992) driving losses in global ice volume (Ullman et

al., 2015). During the deglacial cycle at the end of the Pleistocene atmospheric CO_2 concentrations increased rapidly followed closely by global temperatures (Shakun et al., 2012; Parrenin et al., 2013). Recent synthesis of Holocene paleoclimatic evidence from ice cores, deep sea sediments, and speleothem records suggests that, globally, this period was marked by an initial warming consistent with a continuation of the late Pleistocene deglaciation, followed by a temperature plateau or cooling trend through the modern era (Shakun et al., 2012; Marcott et al., 2013). Evidence from multiple Western North American proxy records suggest a climate warmer and drier than modern period during the early to mid Holocene (Cole, 2009). However, sea surface temperatures (SST) along the Northern California coast may have been cooler in the mid-Holocene (Barron et al., 2003), and diatom records from the northern Gulf of Mexico suggest an amplified North American Monsoon (NAM), wet summer, pattern around the middle of the Holocene (Poore et al., 2003). The Holocene appears to be a period of relatively stable vegetation (Cole, 2009), however some notable shifts have been attributed to minor shifts in climate. For example, decreases in the lower elevational limit of the rosaceous shrub, *Coleogyne ramosissima* of approximately 50-100m in the northern Mojave Desert supports the assertion that much of the Holocene was warmer and/or dryer than the present and that the last few

centuries were anomolously cooler and/or wetter than the rest of the period (Cole and Webb, 1985). This cooler/wetter record, potentially corresponds to the effects of Little Ice Age climate event influencing Western North American climate (Mann et al., 2009).

Packrat Midden Plant Macrofossil Record

Plant macrofossils collected from packrat (*Neotoma* spp.) middens have documented changes in the flora of the American Southwest over the last ~50kya (Wells and Jorgensen, 1964; Wells and Berger, 1967; Martin, 1969; Van Devender, 1977; Van Devender and Spaulding, 1979; Betancourt and Van Devender, 1981; Betancourt et al., 1990). *Neotoma* rats collect plant material for the purpose of building nests and storing food. Nest materials and discarded food fragments left in dry environments, such as caves and under rocky overhangs, can remain intact for thousands of years, deposited in layers representative of times of occupation. Though each species has preferences for specific plants in their diet and nests it has been observed that their foraging is not exclusive to those plants and will result in a representative sample of the local (<50m radius from the midden) flora (Dial and Czaplewski, 1990). Though biased in proportions of plant material collected vs. actual density or abundance in the vegetation, these floristic vegetation records are excellent source of primary taxonomic

characterization of the vegetation. Studies have shown that modern *Neotoma* middens include plant material from 65-85% of the locally occurring species (Dial and Czaplewski, 1990; Finley, 1990). High fidelity taxonomic representation of the vegetation provides good primary data for inferring quantitative climate via the Climate Reconstruction Analysis using Coexistence Likelihood Estimation (CRACLE) methodology (Harbert and Nixon, 2015), particularly since CRACLE is not sensitive to biased sampling as long as samples are sufficiently large in terms of numbers of species represented.

The plant macrofossil record made available by the USGS Packrat Midden Database (geochange.er.usgs.gov/midden/) provides an ideal dataset for the application of CRACLE for paleoclimate estimation. CRACLE will estimate fossil taxon climate tolerances using a Smallest Inclusive Group (SIG) approach (see Chapter 2), which is distinct from the Nearest Living Relative methodology (Mossbrugger and Utescher, 1997). In a SIG approach, extinct species of species that cannot be identified confidently as equivalent to modern species are replaced by surrogates, either based on a phylogenetic analysis or a reasonable taxonomic group such as genus which can be assumed to be a monophyletic group. Most of the fossils documented in the USGS Packrat Midden Database have been identified to a modern genus

or species, and thus, can be either used directly as species or generic surrogates for unknown or unidentifiable species. Using modern species/genus distribution data for fossils that clearly fit within these groups avoids issues of niche shifts and extinction associated with applying the NLR approach to older fossil assemblages (Grimm and Potts, 2015). Genera, at least, are generally expected to adhere to coherent distributions in climate space just as do species (Huntley et al., 1989; Ackerly, 2003; Wake et al., 2009). It has even been suggested that geographic ranges and climate niches are more stable at the genus level owing to intrinsic features of the group being unlikely to change through time (Hadly et al., 2009). CRACLE will be applied assuming relatively stable climate occupancy of these species, and /or surrogates derived from genera over the last 50,000 years. The CRACLE method is expected to maintain accuracy using this SIG approach (Chapter 2; see also Harbert and Nixon, 2015).

Filtered fossil sites for this study were limited to a maximum age of 50,000 years due to carbon dating uncertainties and low sample sizes beyond that time. The dataset included midden fossils as young as just a few hundred years old. The time periods of this study that are best represented are near the terminal Pleistocene glacial cycle (21ka to 10ka – ka ~ *kilo annum* before present), and the middle to late Holocene (<5ka). On average there were about

19 taxa used per sample, and about 40% of these were coded as genera. Average taxon number per site was variable through time (Fig. 3.2). When placed into 200 year bins to generate a centennial scale timeline there were, on average 10 to 20 midden samples per 200 year bin (Fig. 3.2). However, the number of midden samples is correlated with age (younger middens are more common; r=0.70, p=0.00) and potentially with warmer estimated temperatures (r=0.47, p=0.00), though one of these relationships could be due to autocorrelation.

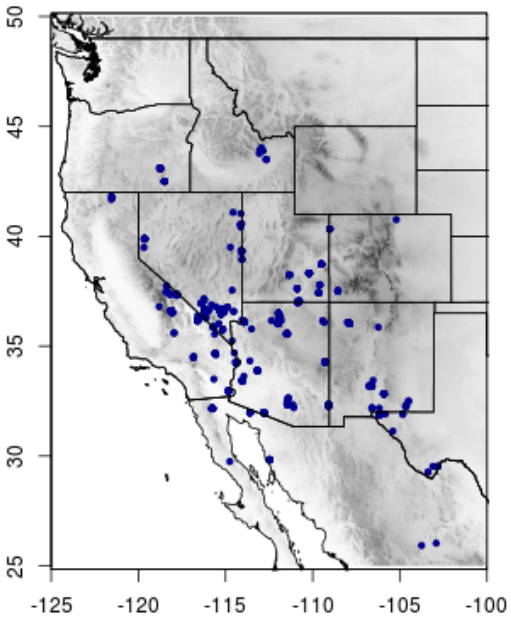

Figure 3.1. Geographic distribution of 179 packrat midden fossil localities used for CRACLE paleoclimate proxy.

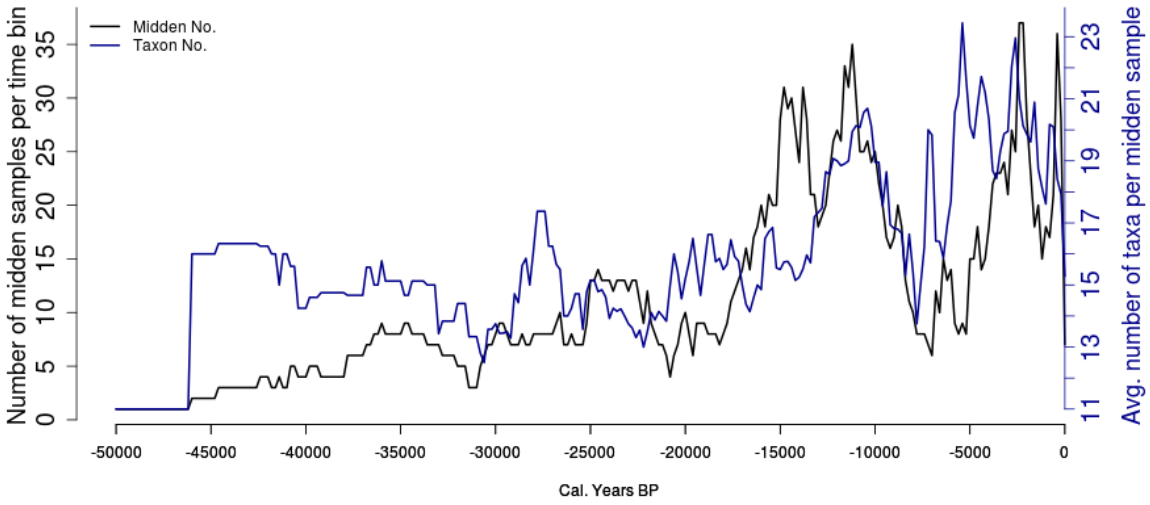

Figure 3.2. Midden representation and average taxon number timeline.
The robustness of the CRACLE model at any time point relies on the number of taxa sampled per site and the number of sites dated to that point.

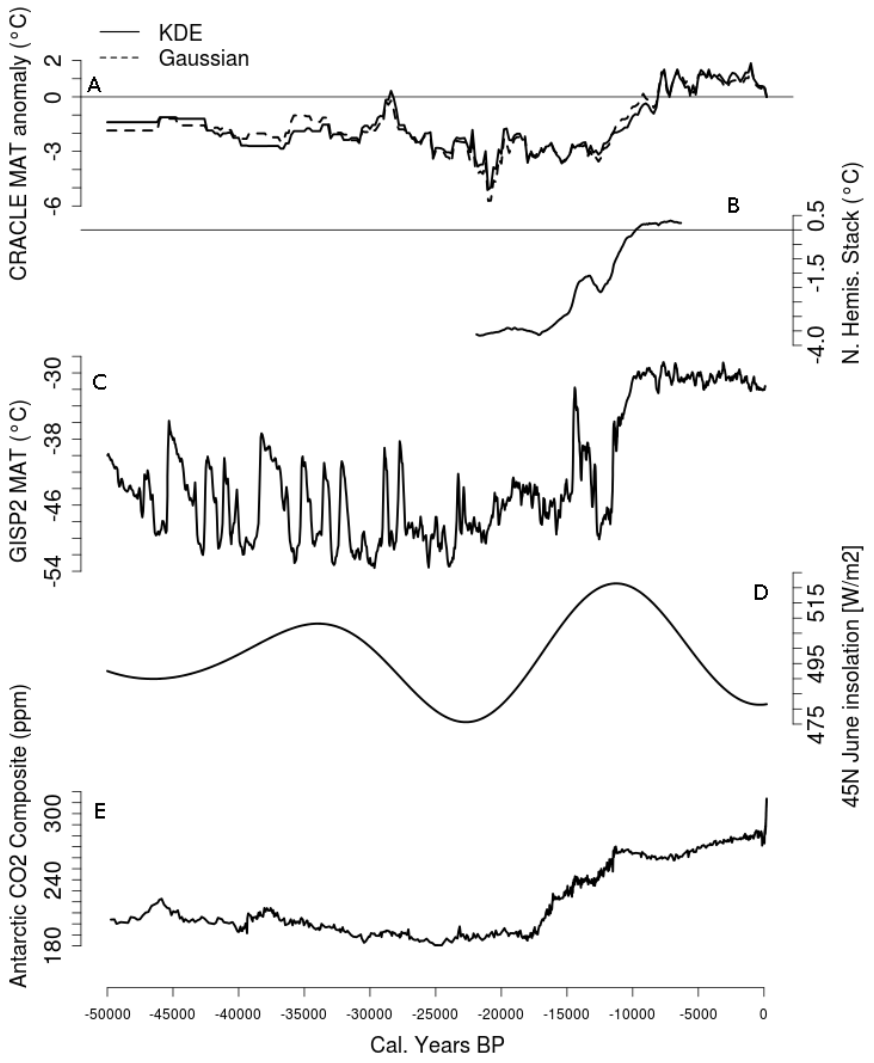

Figure 3.3. 50,000 years of CRACLE estimated Mean Annual Temperature record, peak summer insolation at 45N, Northern Hemisphere Ice-Core estimated temperature data, and atmosphere CO_2 concentration. A) Change in Mean Annual Temperature (MAT) inferred by CRACLE shown as the deviation from the WorldClim 1950-2000 averages for the 693 study sites. B) Northern Hemisphere multiproxy temperature anomaly stack (Shakun et al., 2012). C) Estimated Greenland air temperature using stable isotope data from ice cores covering the last 50,000 years published as part of the Greenland Ice Sheet Project 2 (Cuffey and Clow, 1997; Alley, 2000; Alley, 2004). D) Mean daily summer

(June) insolation at 45N (kJ/m2) estimated following the geometric method described by Berger (1978) using the 'palinsol' R library (Crucifix, 2016). E) Mean CO_2 concentration (ppm) estimated from a composite of Antarctic ice core records (Bereiter, et al. 2015)

CRACLE climate scenario

The CRACLE estimates of mean annual temperature (MAT; Fig. 3.3A) indicates that most of the last glacial period deviated from modern temperatures by -2°C to 4°C, with the exception of a brief colder period reaching to about -6.5°C deviation at 21ka. A punctuated warming trend began after the last glacial maximum (LGM; ~21ka), with three pulses of cooling occurring at about 18.5ka, 16.5ka, and 14ka. Rapid warming began about 12.5ka and continued to the Holocene thermal optimum of +1.5°C by 7ka. The WorldClim 1950 – 2000 MAT average for the midden localities is cooler than most of the MAT time series inferred for the last 9,000 years.

Climatic variables representing seasonal and annual averages of minimum and maximum temperature (Fig 3.4A, 3.4B), precipitation (Fig. 3.4C), available moisture (annual water balance, Fig. 3.4D), and estimated winter length (Fig. 3.4E) were also calculated. These climate estimates characterize the LGM as colder than today by up to -6°C relative to modern annual temperature minimum and maximum values (Fig. 3.4A, 3.4B), and wetter by more than 60 to 100mm (Fig. 3.4C) of annual precipitation (relative

to an average mean annual precipitation across the study localities of 242mm). Prior to the LGM, CRACLE results give evidence that maximum temperatures were depressed by up to -4°C 30-50ka, but minimum temperatures were near modern averages through that same period (Fig. 3.4A, 4B). All CRACLE temperature reconstructions show that the LGM cooling began in this region around 28ka (Fig. 3.3A, 3.4A, 3.4B). Decreases in temperature and increases precipitation during the LGM correspond to a significant increase in the available moisture (Fig. 3.4D). Also, contributing to a more positive estimate of water balance is the inferred increase in winter length by 1 to 3 months during the terminal Pleistocene glaciation (Fig. 3.4E). Near the end of the Pleistocene precipitation begins decreasing (Fig. 3.4C) as mean annual temperatures begin increasing (Fig. 3A),.

Based on CRACLE estimates much of the Holocene is characterized as warmer or at least as warm as the present, as well as dryer (Fig. 3.3A, 3.4). The minimum and maximum temperature estimates indicate that the minimum temperatures responded to a greater degree to the warming deglacial trend than did maximum temperatures (Fig. 3.4A, 3.4B) with up to +3.5°C minimum temperature anomalies through this time and equitable maximum temperature reconstructions. Reconstructions of precipitation and water balance suggest that the Holocene after ~8ka was about as

dry as the current climate.

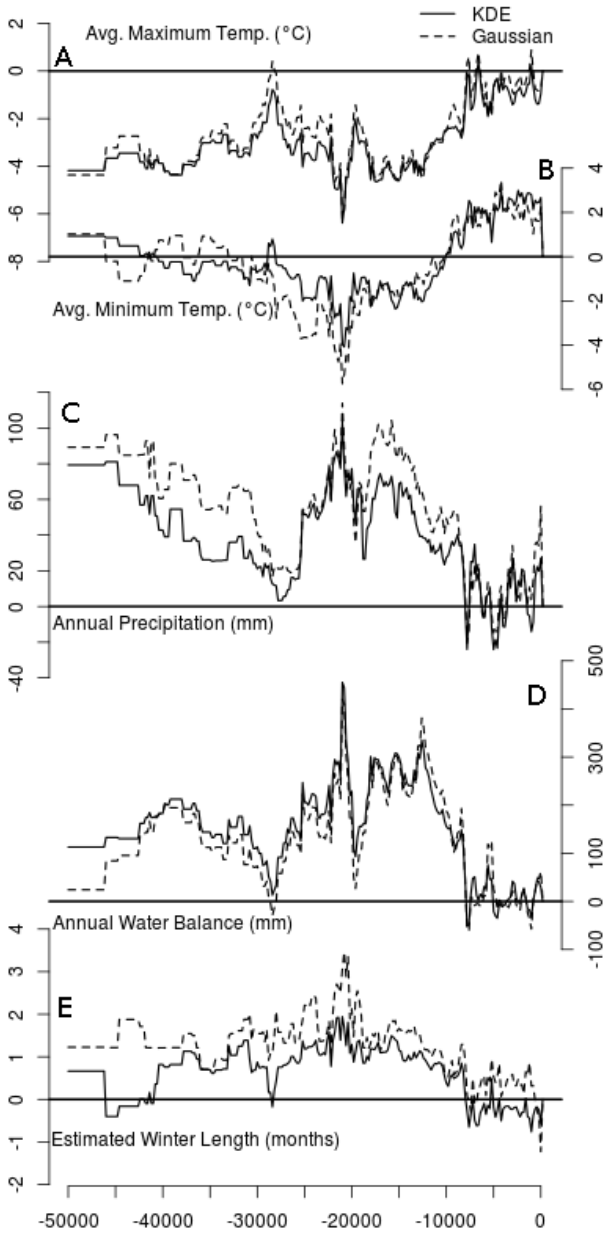

Figure 3.4. Temperature extremes, precipitation, water balance, and winter season CRACLE anomaly time series. A) Maximum annual temperature (°C), B) minimum annual temperature (°C), C) mean annual precipitation, D) mean annual water balance = potential evapotranspiration + precipitation

(Thornwaite, 1948), E) estimated winter length (months with mean temperature less than 5°C). Solid lines indicate kernel density estimator (KDE) CRACLE and dashed line indicate Gaussian distribution CRACLE results.

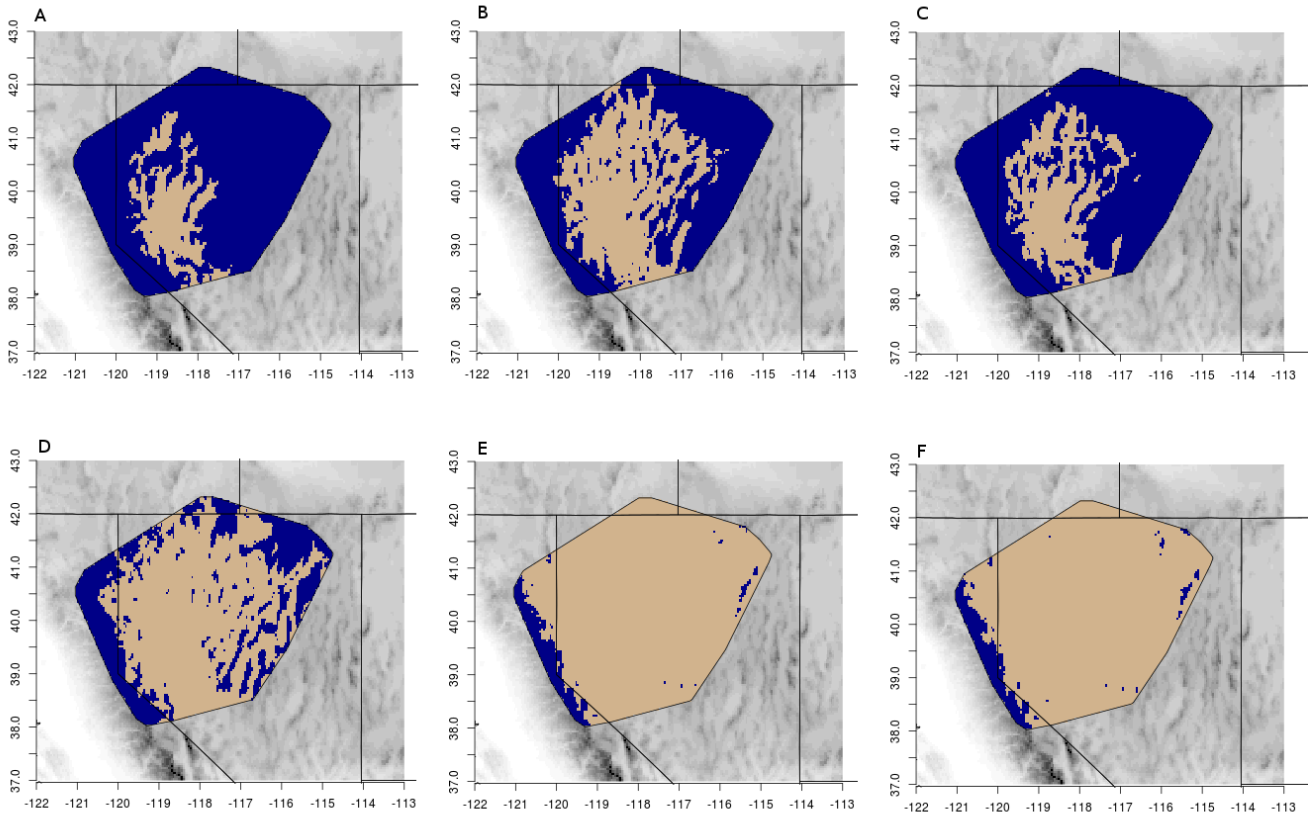

Figure 3.5. Western Great Basin Pleistocene pluvial lake basin spatial analysis of positive/negative water balance reconstructed by CRACLE. Visual representation of the area of the Western Great Basin drainage showing estimates of positive water balance (blue) vs. negative water balance (tan) at A) the LGM – 21ka, B) the pluvial lake level maximum (highstand) – 17ka, C) the Younger Dryas episode – 12.5ka, D) Pleistocene/Holocene transition – 10ka, E) the Holocene Thermal Optimum – 7ka, F) modern – 1950-2000 WorldClim Average (Hijmans et al., 2005).

CRACLE climate context

The ~50,000 year climate profile inferred from the CRACLE estimates is broadly consistent with existing literature on the pattern of northern hemisphere temperature changes over the last ~50,000 years (Cuffey and Clow, 1997; Alley, 2000; Alley, 2004). Coldest temperatures were observed near the last glacial maximum (LGM) around 21ka, followed by a warming trend beginning approximately 15ka and continuing until after the onset of the Holocene (Fig. 3.3), appearing consistent with an increased northern hemisphere insolation forcing (Fig. 3.3D; Berger, 1978), and/or trailing the increased atmospheric CO_2 and resultant greenhouse forcing recorded in the Antarctic ice (Bereiter et al., 2015) consistent with recent investigations into the deglacial warming – CO_2 relationship (Shakun et al., 2012; Parrenin et al., 2013).

Based on external data, the deglacial warming was interrupted with slight cooling episode beginning after 14ka and continuing until 12ka (Shakun et al., 2012), resulting in a CRACLE estimated decrease in MAT by ~0.5°C. This climatic event could correspond to the vegetational response to the Younger Dryas episode observed in the Grand Canyon, Arizona (Cole and Arundel, 2005). This was followed by largely uninterrupted warming until the Holocene Thermal Optimum at 7.5ka (Fig. 3.3A). The total

warming over this time period (~5ky) estimated by CRACLE across Western North America was less than 5°C.

The wetter glacial pattern, however, remained largely intact according to the CRACLE estimates until after the Pleistocene-Holocene transition (Fig. 3.4C, 4D). This wetter pattern is likely due to a southern displacement of the circumpolar jet stream by the Laurentide ice sheet and therefore increased winter storm frequency and precipitation across the study area (Oster et al., 2015). Evidence for limited reduction in ice sheet surface area, with volume losses instead due to reduction in height (Ullman et al., 2015), suggest that the influence of the ice sheet into the early Holocene may be reflected in the CRACLE moisture reconstructions (Fig. 3.4C, 4D). Interestingly, estimated winter length remained longer than the modern until around this same time (Fig. 3.4E), despite both rising annual (Fig. 3.3A) and minimum temperatures (Fig. 3.4B). The rise of the Pleistocene pluvial lakes in closed basins across the region are attributed to this increased influx of moisture and cooler temperatures (MacDonald et al., 2008; Lyle et al., 2012; Ibarra et al. 2014; Lachniet et al., 2014; Oster et al., 2014), this is in contrast to the dry-cold glacial hypothesis (Brackenridge, 1978).

In order to visualize the change in area of positive water balance, the CRACLE reconstructions of water balance were applied across the former Lake Lahontan basin (Reheis, 1999) in

the western Great Basin at time points through the glacial period and Holocene (Fig. 3.5). These results indicate that during the glacial period there was a large increase in areas of positive water balance across the Lake Lahontan prior to the Pleistocene-Holocene transition (up to 83% at the LGM – 21ka – vs. just 7% in the WorldClim 1950-2000 modern average). This would have corresponded to a massive influx of surplus water into the basin and contributed to the increased lake levels. The CRACLE reconstructions of temperature and precipitation indicate this increase in water balance through the glacial period results from both decreases in temperatures (-3-6°C from modern) and increases in precipitation (+60 to 100mm) (Fig. 3.3A, 4).

Independent multiproxy climate reconstructions of the Holocene paleoclimate suggests this period was marked by an initial warming, consistent with a continuation of the late Pleistocene deglaciation, followed by steady temperatures or a slight cooling trend through the modern era (Marcott et al., 2013). The amplitude of these changes is shown to be higher at high latitudes (Marcott et al., 2013). These global proxies are generally consistent with the results of these CRACLE analyses with the exception that CRACLE estimates indicate a higher amplitude temperature shift (up to +2°C) during the Holocene (Fig. 3.3A), and a mostly steady climate through the Holocene. Alternatively, a

recent re-examination of global temperatures and atmospheric composition through the Holocene has promoted hypotheses that early human activities, especially agriculture (Fuller et al., 2011) and deforestation (Fyfe et al., 2015; Li et al., 2009), may have contributed to an anomalous increase in atmospheric concentrations of CO_2 and CH_4 and subsequently changed the dominant climatic forcing factor from insolation to greenhouse gases (Ruddimann et al., 2015). The CRACLE climate reconstruction shows no steady downward trend in MAT between the Holocene Thermal Optimum (HTO) and the end of the Holocene (Fig. 3.3A). However, CRACLE results do show that the final millennium of the time-line features a downward trend of about 1.5°C (Fig. 3.3A), consistent with fossil evidence of changes in vegetation (e.g., Cole and Webb, 1985) and dendrochronology estimates of paleotemperature (Briffa, 2000). Complicating the interpretation of the Holocene climate, CRACLE results indicate a slight increase in rainfall after the HTO (Fig. 3.4C), which would have offset the drying effects of the warmer temperatures. The increase in rainfall is consistent with increasing influence of the North American Monsoon (NAM) across the west (Poore et al., 2005).

Conclusion

The late Pleistocene plant fossil record preserved in packrat midden macrofossil deposits provides appropriate data for CRACLE estimation of a complete, well supported, timeline of climatic change across western North America. This is the first application of CRACLE to paleoclimate estimation and the first quantitative reconstruction of climate in this region using plant macrofossils. The detailed climate profile estimated using CRACLE includes many well-understood features of global climate based on other data sources and methods of analysis. The cold LGM and Younger Dryas are prominent. Temperature increases associated with global deglaciation in the study area are preceded by increased atmospheric CO_2 concentrations, providing a possible analog for the studying climatic and biotic change similar to the modern climate change scenarios. Perhaps more importantly, this reconstruction quantifies changes in numerous climate variables not typically accessible through other proxy methods (e.g., water balance and winter length) and provides much higher precision for such variables than can be attained with other methodologies. The climate profile also sheds light on key events of the geologic and fossil record including the development of large pluvial lakes in the Great Basin during the terminal Pleistocene glacial cycle. The rapid temperature increase and precipitation decrease marking the

Pleistocene-Holocene transition is evidence of rapid climate change that elsewhere has been implicated in coordinated megafaunal extinctions of mammoths and horses in North America (Guthrie, 2006). These results illustrate that CRACLE can provide robust, high-quality estimates of climate in recent geologic time when appropriate high quality samples are available, and that vegetation responds rapidly (on a geologic scale) to climatic changes.

CHAPTER 4

CLIMATE NICHE MODELING IN THE PERENNIAL *GLYCINE* (LEGUMINOSAE) ALLOPOLYPLOID COMPLEX

Previously published as:

Harbert, R.S., A.H.D. Brown, and J. Doyle. 2014. Climate Niche Modeling in the Perennial *Glycine* (Leguminosae) Allopolyploid Complex. *American Journal of Botany* 101(4):710-721.

Reprinted here with permission from the American Journal of Botany. Pagination here differs from the original.

Abstract

Premise of study: Polyploid plants, when compared with diploids, show similar molecular, morphological, physiological, and ecological tendencies across unrelated groups, but the degree to which these form “rules” of polyploid evolution are unclear. The *Glycine* (Leguminosae) allopolyploid complex affords the opportunity to test whether polyploidy in similar genetic backgrounds produces similar effects on geographical range or

climatic space.

Methods: We used information on locality presence of four closely related *Glycine* allopolyploid species and their diploid progenitors to build models of the potentially available Australian ranges based on climate using Maxent3.3.3k. Principal coordinate analysis was used to characterize the multidimensional climate space occupied by each species.

Key results: Each of the four *Glycine* allopolyploids showed intermediacy in potential geographical space and in ecological space, relative to its diploid progenitors. The four allopolyploids did not have consistently larger ranges than their progenitors, though all four occupied a portion of climate niche space not available to its progenitors. The polyploids also differed in their exploitation of potentially available geographical range. Australian ranges and environmental space did not correlate with greater colonizing ability in these polyploids.

Conclusions: The four *Glycine* allopolyploids do not show many common range or climate related features, other than intermediacy. Thus, despite their similar genetic and evolutionary backgrounds, polyploidy has not produced convergent ecological effects.

Introduction

Polyploidy (whole genome duplication) is a key genetic phenomenon in plant evolution. All seed plants are descended from an ancestor that had experienced a whole genome duplication, and the genomes of all extant angiosperms have been shaped by one or more additional polyploidy events (Soltis et al., 2009; Jiao et al., 2011). Approximately 15% of all speciation events in flowering plants are estimated to have involved polyploidy, and 35% of all extant angiosperms are chromosomally polyploid (Wood et al., 2009). Why polyploidy is so prevalent has long been discussed. At one extreme, it has been suggested that no adaptive advantage need be postulated and that polyploids simply may accumulate as part of a one-way "ratchet" process in which diploids can give rise to polyploids, but polyploids never give rise to diploids (Meyers and Levin, 2006). On the other hand, it has often been hypothesized that polyploids have attributes that make them more adaptable, better colonizers, and even invasive relative to diploids (e.g., Otto and Whitton, 2000; Pandit et al., 2011; Madlung, 2013).

A major theme in research on polyploids is the search for "rules" (Soltis et al., 2010)—are there emergent properties that characterize polyploids as a group and that transcend taxonomy? Polyploidy itself is a convergent feature in plant evolution; but what

attributes do independently formed polyploids share? In all taxa, whole genome duplication in the short term increases genome size of a polyploid relative to its progenitor(s). Increased DNA content has predictable “nucleotypic” (Bennett, 1972) effects regardless of ploidy, for example, leading to increases in cell cycle time (Francis et al., 2008) and cell size (Beaulieu et al., 2009). Polyploids may undergo convergent longer-term changes as part of the process of diploidization (e.g., Doyle et al., 2008), including such phenomena as genome size reduction (Leitch and Bennett, 2004) and the loss of sequences from duplicated regions (“fractionation”: Freeling et al., 2012). Gene loss may follow predictable patterns (e.g., Blanc and Wolfe, 2004; Paterson et al., 2006; Birchler and Veitia, 2014), but there are also differences among lineages (Barker et al., 2008).

Similarly, some of the many effects of polyploidy on morphology, physiology, and ecology (reviewed by te Beest et al., 2012) show common trends across unrelated taxa. For example, Warner and Edwards (1993) reported that polyploids generally have higher photosynthetic rates per cell than their diploid relatives. Perhaps as a result of physiological or biochemical attributes (e.g., Pearse et al., 2006), polyploids are often invasive (Pandit et al., 2011; te Beest et al., 2012). The effect of polyploidy on range has been a topic of long-standing interest in the polyploidy literature, linked to the debate concerning the ecological success of polyploids

relative to diploids (e.g., Madlung, 2013). The fact that the Arctic flora contains a high percentage of polyploids supports the notion that polyploids can exploit newly available habitat (Brochmann et al., 2004). But as Ramsey (2011) has noted, it is difficult if not impossible to know whether such a correlation is due to polyploidy, per se, or is the product of thousands to millions of years of adaptive evolution in species that happen to be polyploid. A recent survey of over 400 species of diploids and related polyploids showed that there was no significant correlation between range or ecological attributes and ploidy (Martin and Husband, 2009). Some earlier studies similarly contradicted commonly held views that polyploids generally have larger ranges (Stebbins and Dawe, 1987; Petit and Thompson, 1999). On the other hand, Lowry and Lester (2006) found that polyploid species of *Clarkia* have larger ranges than their diploid progenitors, possibly because they are better colonizers. Hijmans et al. (2007) found that among 185 *Solanum* species, diploids and polyploids had similar range sizes, but that the two species with the largest ranges were both polyploids; they suggested that polyploidy had led to an expansion of the group into previously unoccupied ecological and geographical ranges. In *Aegilops*, Meimberg et al. (2009) found a correlation between multiple origins of polyploids—leading to increased genetic diversity—and increased geographic range.

Recently, climate niche modeling has been used to investigate whether the niches of polyploids differ from those of their close diploid relatives. Oberprieler et al. (2012) studied 20 Iberian taxa of *Leucanthemum* (Asteraceae), ranging in ploidy from 2 x to 22 x; they found that although the most widespread species was a hexaploid, and in general the largest ranges were those of tetraploids, there was a negative correlation overall between range size and ploidy. They found no correlation between ploidy and eco-climatologically modeled potential range. Theodoridis et al. (2013) studied four species of the *Primula* sect. *Aleuritia* (Primulaceae) polyploid complex, comprising a diploid, a tetraploid, an allohexaploid formed from the 2 x and 4 x species, and an allooctoploid formed by a cross between the 2 x and 6 x species. They found that climate niches of these species differed from one another, with polyploids unexpectedly occupying narrower geographical and environmental niche spaces than the diploid. In the taxonomically difficult *Claytonia perfoliata* polyploid complex (Portulacaceae), where exact progenitor relationships remain unclear, McIntyre (2012) found variable levels of niche differentiation between diploids, tetraploids, and hexaploids, with no over-all correlation between niche size and ploidy. Godsoe et al. (2013) concluded that climate niche differentiation did not explain range differences between diploid and autopolyploid cytotypes of

Heuchera cylindrica (Saxifragaceae). The relationship between polyploidy and range or climate niche size is inconsistent and appears to be taxon-specific.

The perennial *Glycine* subgenus *Glycine* (Leguminosae, Phaseoleae) allopolyploid complex is a good model for studying whether allopolyploidy has a consistent effect on range. The complex is composed of eight allopolyploid (2 n = 78, 80) species formed from different combinations of eight extant diploid (2 n = 38, 40) progenitor species. All eight allopolyploids appear to have more recent origins than the divergence times of diploid species in the subgenus and share nuclear alleles and chloroplast haplotypes with their diploid progenitors (Doyle et al., 2004; Bombarely et al., 2014; Fig. 4.1). All of the approximately 26 diploid species in the subgenus are native to Australia, with one diploid species also having populations on Papua New Guinea. The eight allopolyploids all have Australian ranges but, in contrast to diploids, five also have populations outside of the range of any diploid species, in West Timor (*G. tomentella* T3; also found in eastern Papua New Guinea), Taiwan and adjacent mainland China (*G. dolichocarpa* Tateishi & H. Ohashi T2, *G. tomentella* T4, *G. pescadrensis* Hayata), the Philippines (*G. tomentella* T4), the Ryukyu Islands of Japan [*G. tabacina* (Labill.) Benth., *G. pescadrensis*], and various islands of the South Pacific (New Caledonia, Tonga, Fiji, Vanuatu: *G.*

tabacina). Thus, these *Glycine* allopolyploids appear to be better colonizers than their diploid progenitors. Here we compare Australian ranges of four *Glycine* triads (an allopolyploid and its two diploid progenitors, following terminology in Theodoridis et al., 2013), and model potentially available geographical ranges and environmental spaces for each species. We look for consistent patterns across the four triads and ask whether attributes of the Australian ranges predict wider colonizing ability in some of the polyploids.

Materials and Methods

Taxon choice and sampling— Four of the eight *Glycine* allopolyploid (designated by "T") species were chosen for study: *G. tomentella* T1, *G. dolichocarpa* (= *G. tomentella* T2, henceforth referred to as " *G. dolichocarpa* T2"), *G. tomentella* T3, and *G. pescadrensis*. The diploid (designated by "D") progenitors of each of the four polyploids were also included: *G. tomentella* D1/D2, *G. tomentella* D3, *G. syndetika* (= *G. tomentella* D4, henceforth referred to as " *G. syndetika* D4"), *G. tomentella* D5A, and *G. stenophita* B.E. Pfeil & Tindale. Thus, a total of four overlapping triads were studied (Fig. 4.1).

The remaining four allopolyploid species were excluded for one of two reasons. *Glycine tomentella* T5 and T6 are known from

only one (T6), or a few (T5) populations, all in Australia, too few for our purposes. The species *G. tomentella* T4 and *G. tabacina* were both formed by multiple origins involving different genotypes from several diploid species (Doyle et al., 1999, 2002), making it difficult to compare polyploid ranges with those of diploid progenitors. In common with the four species under study, both of them are transoceanic migrants.

The genus *Glycine* includes the cultivated soybean (*G. max*) and therefore has been the target of extensive collections in Australia. The CSIRO (Australia) Perennial *Glycine* Germplasm Collection includes seed from over 2000 accessions representing all of the species in subgenus *Glycine*, each with longitude and latitude information. These accession coordinates generally have significant digits to one or two decimal places and therefore are precise to a few kilometers in most cases. This precision is generally sufficient given the 2.5 arcmin (0.042 °) resolution of the climate data used for this study (Hijmans et al., 2005). Geographic information from all available accessions of each relevant species was used, with the exception of a few accessions where the record for the location of origin is highly questionable (A. H. D. Brown, unpublished data), and one outlier population of *G. tomentella* D3 (G2586, 1729 km from the nearest other *G. tomentella* D3 population). This database was further trimmed so that for each species only locations in

Australia and Papua New Guinea were considered for model building. This is a reasonable assumption about the “primary” range of these *Glycine* species, because all of the allopolyploids are relatively recent derivatives from diploids that are confined to these regions (Doyle et al., 2004). Presence records were only considered once for each locality (exact latitude and longitude) in an effort to minimize any collection or spatial bias, since multiple collections at a single site may not actually indicate prevalence in that area. This database was also used to determine target-group absence locations (Phillips et al., 2009; Mateo et al., 2010), as discussed below. The number of unique accession localities used per species was: *G. tomentella* D1 (27), *G. tomentella* D3 (44), *G. syndetika* D4 (11), *G. tomentella* D5A (29), *G. stenophita* (34), *G. tomentella* T1 (138), *G. dolichocarpa* T2 (23), *G. tomentella* T3 (11), and *G. pescadrensis* (79) (Appendix S4.1).

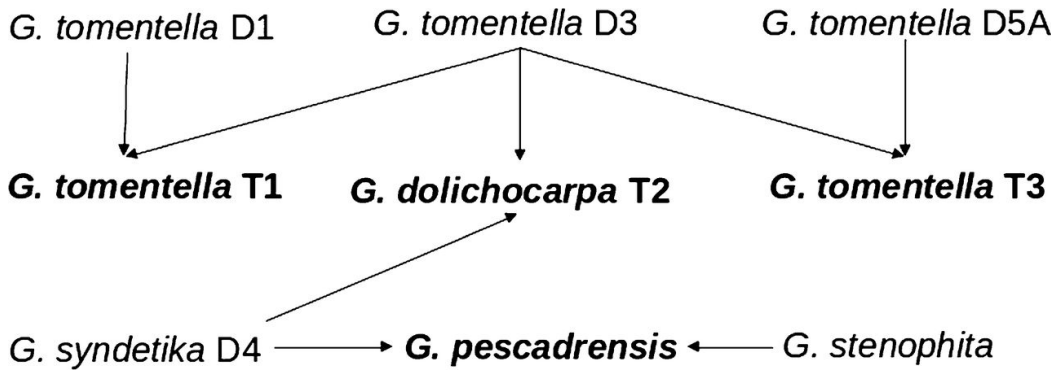

Fig. 4.1. Relationships among four allopolyploid Glycine species (in bold) and the five diploid species that contributed their homoeologous genomes. (Doyle et al., 2004).

Climate data— The 19 climatic variables used for this study came from the Worldclim database (http://www.worldclim.org). This database was generated by compiling data from multiple climate databases and nearly 50 000 weather stations worldwide. A 0.5 arcmin grid was assembled using these data to interpolate climate parameter values at each grid cell (Hijmans et al., 2005). These 0.5 arcmin grids were used to generate lower resolution grids at 2.5, 5.0, and 10 arcmin. We chose to use all 19 (Appendix S2, see online Supplemental Data) of the bioclimatic variables at a resolution of 2.5 arcmin to incorporate climate heterogeneity between species occurrence localities. The lower resolution grid (2.5 arcmin, about 4 km at the equator) was chosen for easier processing and to match

best the precision of the occurrence data.

Suitable climate niche modeling— Maxent — Climate-based species distribution modeling (SDM) by the maximum entropy algorithm—Maxent—was performed using the R software package 'dismo' (R Development Core Team, 2011; Hijmans et al., 2012) and the default settings of Maxent 3.3.3k (Phillips et al., 2004; Phillips and Dudik, 2008; Elith et al., 2011). The Maxent protocol uses presence only data and does not assume that occurrence data represent prevalence of a species at each site (Phillips and Dudik, 2008). The principle of maximum entropy is used to generate a set of functions that assign a value scaled to be between 0 and 1 corresponding to the suitability of each available climate for that species where 0 is not suitable, and scores generally above 0.1 indicate varying degrees of suitability, increasing as the score gets larger (Phillips et al., 2006). The model fitted by Maxent can then be used to estimate the potentially available geographical space for a species based on climate suitability (Phillips et al., 2006).

Warren (2012) argued that analyses like these should be termed “niche models” rather than “species distribution models” because of their underlying assumption that the model is describing some element of the niche for that species. Due to limitations of available predictor and occurrence data, these models may not

accurately describe the true constraints on a species' distribution—the niche (Warren, 2012). We acknowledge that the models presented here may be imperfect representations of the niche of each species, but they are useful nonetheless because they provide a framework that applies a uniform set of methods to correlating a species' distribution with climatic factors that is relatively robust with small sample sizes (Hernandez et al., 2006; Phillips and Dudik, 2008).

Model evaluation— Negative occurrence data for model evaluation were generated employing the principle of "target-group absence" (Mateo et al., 2010). Each locality of *Glycine* occurrence in the study area was evaluated to determine whether those coordinates were at least the width of two entire climate grid cells (5 arcmin ~0.042°) from the nearest occurrence of a species that does not occur at the starting locality. If these criteria were satisfied, then the coordinate was added to a list of implied absences for the target-group. The presence records for *Glycine* were obtained as part of a directed effort to collect *Glycine* specimens for germplasm, and therefore at any location, all *Glycine* present were collected. Gonzalez-Orozco et al. (2012) noted the contrast between such a collection strategy and a more haphazard general botanical collection. Genus-wide directed collecting should

improve the reliability of implied absence records generated through the target-group absence theory.

Model evaluation using receiver operator characteristic (ROC) curves was implemented for the Maxent models generated by this study. The area under the receiver (or relative) operator characteristic curve, known as the AUC value, is a widely used test of predictive ability of species distribution models (Liu et al., 2011). The AUC is a measure of the rate at which the model predicts known presence localities over known or inferred absence locations when presence and absence localities are chosen at random. A random model would not favor presence sites over absence sites and would therefore predict false positives and false negatives at the same rate (AUC = 0.5), whereas a perfect model would predict no false positives or false negatives AUC = 1 (DeLong et al., 1988; Phillips et al., 2009; Elith et al., 2006). In evaluating models for this study, we employed a bootstrap resampling of presence and target-group absence data to reduce the effects of spatial bias in the samples of points used to generate the ROC curves. Each model was evaluated 1000 times, each time using a random subset of 100 target-group absence records and 80% of the presence records.

Geographical space (G-space) comparisons— The Lowest Presence Threshold (LPT) technique (Pearson et al., 2007) was

employed here to make the distribution models easier to visualize and for comparison of potential and realized geographical space (G-space). This technique sets the threshold of positive prediction of a locality as belonging to the range of suitable climate (the potential G-space) at the lowest value assigned to a known presence locality for the species being modeled. Application of the LPT values ensures that the omission rate of positively occupied locations is zero and the widest confirmed suitable climatic space is mapped as the potential range. The LPT can vary between models, but by using this method the ranges that are predicted are consistent for all of the models. The LPT was used to assign binary (true/false) designations to known Glycine presence localities outside of Australia as a metric of whether generally suitable climate exists for a species regardless of whether that species occurs at that locality. This allows examination of whether the polyploids have dispersed as a result of expansion in their suitable climate niche over that of their related diploids.

The potential ranges reported here are estimates of area in square kilometers calculated from the number of 2.5 arcmin climate grid cells predicted as suitable (Maxent score > LPT) with cell area adjusted for the latitude. This method of area estimation is slightly biased because it calculates the width of each cell in kilometers at the middle and treats it as a rectangle to generate the area. This is

a more accurate estimate of range size than the cell count alone (Hijmans and Van Etten, 2012). Realized ranges were calculated via a modified minimum convex hull (MCH) method. The MCH is the smallest convex polygon that can be drawn around a cluster of points. The MCH was calculated using the accession locality coordinates for each species and then modified by overlaying onto the map of LPT limited potential G-space to only include the area within this polygon that was also ranked at or above the LPT by the Maxent models. This results in a subset of the potential range that is an appropriate approximation of the area that is actually occupied by the species and can be considered to be the realized range for the purposes of this analysis. The method used to estimate area of potential ranges was applied for the realized ranges.

Geographical space potentially and actually occupied by each species was compared within triads. Schoener's D (dissimilarity) index (Schoener, 1968), an accepted method for the comparison of species ranges (Warren et al., 2008), was calculated for each pairwise comparison within triads of climatically suitable G-space. Schoener's D was calculated from raw Maxent scores as well as from binary suitability predictions based on the LPT. Schoener's D calculated for the binary suitability predictions is not influenced by specific Maxent scores and therefore may be a less-biased

representation of the difference in G-space (Warren et al., 2008). Null distributions of D values for raw Maxent scores and the LPT-based predictions were calculated as described in Warren et al. (2008) by iterative model construction using two random samples of occurrence records from the pool of all occurrence localities in the study area. These null distributions provide a test of significance for pairwise niche similarity comparisons via Schoener's D index. The area of geographic overlap between members of a triad was analyzed both for potential and for actual ranges.

All range size calculations were limited to continental Australia to standardize the study area. The realized and potential ranges were compared to determine the percentage of the theoretical range that is occupied by the species and the pairwise overlap between the study species potential and realized ranges within triads.

Environmental space (E-space) comparisons— The climate niche was visualized independent of geography using principal coordinate analysis (PCoA) with software from the R Project for Statistical Computing (R Development Core Team, 2011; Oksagen et al., 2012). The centroids of species distributions in this E-Space were then plotted and the scatter of the distribution used as a measure of variance around the centroid, allowing for significance

testing of centroid position shifts between species in each triad. Differences in niche size between species were measured by comparing the average distance to the centroid within a triad. In all relevant pairwise comparisons (within triads as well as using all taxa), there were unequal sample sizes and unequal variances. Because this data set has a limited number of samples (accession localities), we applied the Games–Howell multiple comparison test (Games and Howell, 1976; Day and Quinn, 1989) to estimate significance of centroid position differences, and Tukey's honestly significant difference test (Tukey, 1953; Hayter, 1984) to compare E-space breadth as measured by the average distance to cluster centroid for each species.

Results

Model evaluation— Model performance varied slightly among taxa but in all cases was adequate based on bootstrapped AUC values (Table 4.1), all of which far exceeded the AUC value of 0.7 recommended as the threshold of satisfactory model performance by Elith et al. (2006).

Comparison of polyploid and diploid actual ranges— Potential ranges were estimated from Maxent predictions of climatic suitability across Australia (Fig. 4.2). The realized geographical

space, or rG-space, comprised the subset of these ranges from where collections for the species were made. The rG-space in the four triads showed no consistent correlation with ploidy (Fig. 4.3). Two polyploids (*G. pescadrensis* and *G. tomentella* T1) have larger ranges than either of their progenitors, but *G. dolichocarpa* T2 and *G. tomentella* T3 have considerably smaller ranges than the more widespread of their diploid progenitors. As a group, polyploids do not have significantly larger actual ranges than diploids (Wilcoxon signed rank = 7, P = 0.34).

In all four cases, a considerable proportion of the range (46–50%, Fig. 4.3) of the polyploids lay outside the combined ranges of their diploid progenitors. Only *G. pescadrensis* had a range that encompassed the combined ranges of both its diploid progenitors; the other three polyploid species occupied considerably smaller (4–65%, Fig. 4.3) portions of the actual ranges of their diploid progenitors.

Table 4.1: Species distribution model evaluation

	AUC[1] Value:			
Model for taxon:	**Mean**	**Min**	**Max**	**# taxon records used**
*G. pescadrensis**	0.90	0.82	0.96	34
G. tomentella T1*	0.96	0.90	1.0	138
G. dolichocarpa T2*	0.99	0.95	1.0	23
G. tomentella T3*	0.98	0.94	1.0	11
G. tomentella D1	0.89	0.78	0.97	27
G. tomentella D3	0.99	0.94	1.0	44
G. syndetika D4	0.96	0.88	1.0	11
G. tomentella D5A	0.98	0.93	1.0	29
G. stenophita	0.93	0.87	1.0	79

[1]AUC – Area under the receiver-operator curve iteratively calculated on 1000 random samples of each species occurrence data and background data.
* Indicates polyploid species.

The zone of sympatry between diploid progenitors is a potential site of origin for a diploid or allopolyploid hybrid and thus could represent a climate niche that is accessible to the product of hybridization. Diploid progenitor pairs for two polyploids (*G. tomentella T3, G. pescadrensis*) had no range overlap (Fig. 4.3). In the other two triads, diploids had very small zones of sympatry (7–8% of their combined ranges), and polyploid species occupied 81% (*G. dolichocarpa* T2) and 100% (*G. tomentella* T1) of these small regions.

Comparison of polyploid and diploid potential ranges— Many factors other than climate, notably dispersal and competition, have shaped the realized ranges of species (Peterson et al., 2011). In contrast, the potential geographical space (pG-space) based solely on climate variables may represent a better estimate of the underlying physiological characters related to climate niche than rG-space. The Maxent predictions for the four tetraploids and their diploid progenitors identified the potentially available ranges for each species (Figs. 1.2, 1.3). Potential and realized ranges were positively correlated (Spearman's rank correlation, $\rho = 0.8$, $P = 0.02$). As with realized ranges, there was no correlation between potential range size and ploidy within triads: only *G. pescadrensis* had a substantially larger pG-space than both of its diploid

progenitors, and both *G. tomentella* T1 and *G. tomentella* T3 had considerably smaller pG-spaces than one of their progenitors. The pG-spaces of the four polyploid species as a group were not significantly larger than those of the diploids (Wilcoxon signed rank = 7, P = 0.36).

Schoener's index (D) of niche dissimilarity (Warren et al., 2008; Schoener, 1968), calculated from raw Maxent scores, showed that in all triads the diploid range models differed more from one another than either one did from the polyploid; thus, each polyploid appears intermediate within its triad (Table 4.2). This observation also held for three of the four triads when Schoener's D was calculated (Table 4.2) for the overlap of pG-space defined by the discrete LPT predictions of range (Fig. 4.2), ignoring the geographic patterns of Maxent suitability scores. The exception was the *G. pescadrensis* triad, where Schoener's D indicated that the tetraploid, *G. pescadrensis*, was more distinct from one diploid progenitor (*G. stenophita*) than that diploid was from the other progenitor (*G. syndetika* D4) when comparing discrete LPT range predictions.

Allopolyploids combine the genomes of their two diploid progenitors, and thus an allopolyploid at its formation should inherit genic variation affecting physiological determinants of geographical range from both parents. Different models of polyploid

inheritance of physiologically relevant traits lead to different predictions concerning the degree to which portions of progenitor diploid ranges should be accessible to derived allopolyploids, or could extend beyond those of its progenitors. We therefore measured the overlap of each polyploid's pG-space with the following portions of its diploid progenitors' potential ranges: (1) the intersection of the pG-spaces of the two diploids (the zone of potential progenitor sympatry); (2) the union of the pG-spaces of the two diploid progenitors (combined potential diploid ranges); and (3) the pG-space of each of the progenitors individually (parental pG-spaces).

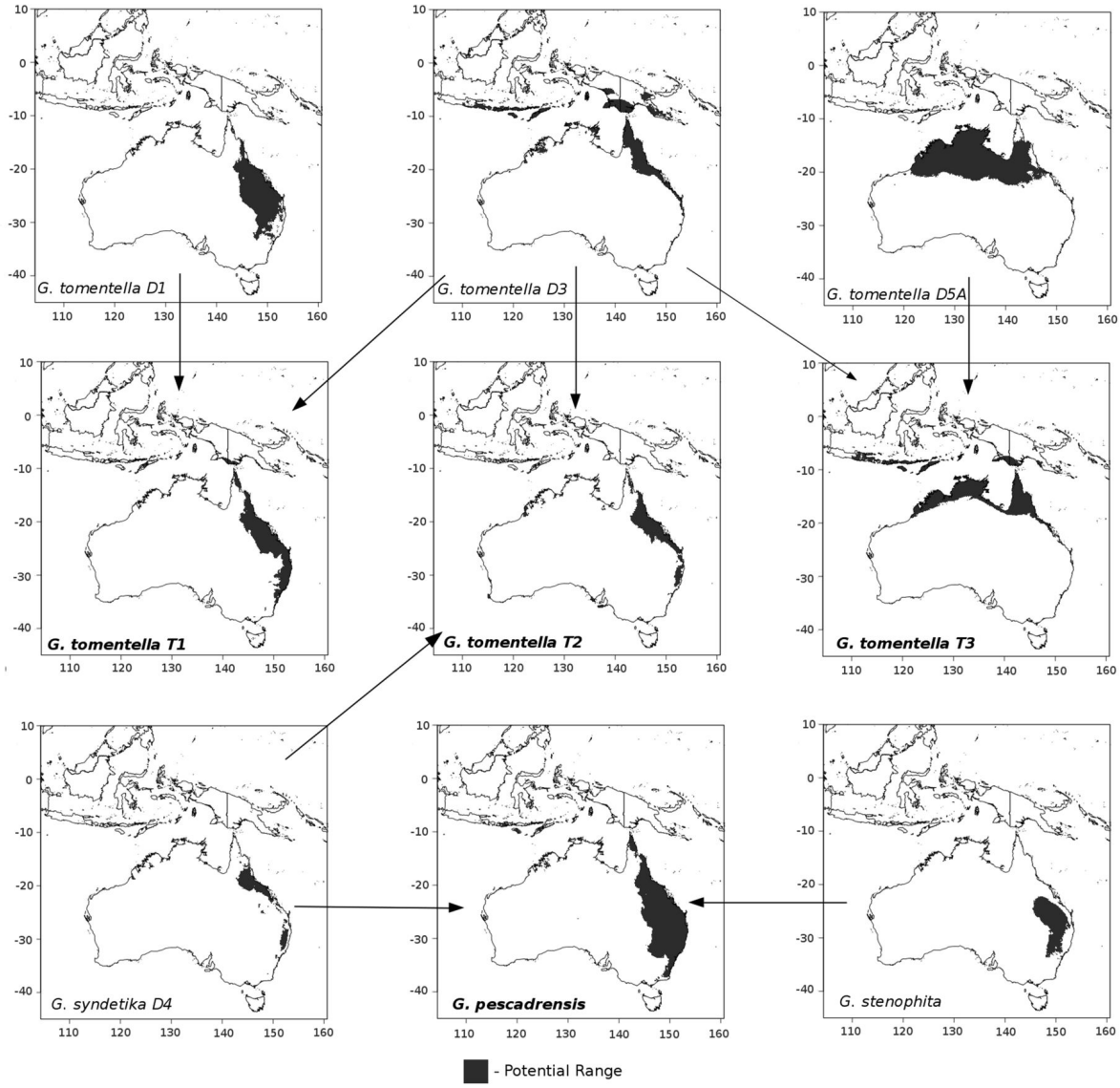

Fig. 4.2. Maxent species distribution models (SDM). The Maximum Entropy (Maxent) algorithm as implemented in R was used to obtain geographic distribution models of suitable climate based on the 19 bioclim variables (online Appendix S2; http://www.worldclim.org). The ranges indicated on these maps are based on the Lowest Presence Threshold (LPT) for each species. Arrows indicate diploid–polyploid relationships as in Fig. 4.1.

The zone of pG-space overlap between pairs of diploid pro-

genitors as a fraction of their combined ranges varied greatly, from 1.4% between *G. syndetika* D4 and *G. stenophita* (*G. pescadrensis* triad) to 36% between *G. syndetika* D4 and *G. tomentella* D3 (T2 triad). *G. tomentella* T1, *G. dolichocarpa* T2, and *G. pescadresis* had pG-spaces that encompassed all, or nearly all, of the zone of overlap between progenitors' pG-spaces, with *G. tomentella* T3 lower but still high (72%).

The pG-space of G. pescadrensis included over 95% percent of the full individual and combined ranges of both parental species. In contrast, the other three allopolyploids varied considerably in this regard (Table 4.3). The pG-space of *G. tomentella* T1 included relatively similar amounts of the pG-space of its two parents (59 vs. 67% of their pG-spaces), whereas *G. dolichocarpa* T2 and *G. tomentella* T3 overlapped considerably more with one parent's pG-space than with the other parent's. Percentage potential occupancy, by the polyploid, of the combined potential ranges of its diploid progenitor also varied in the remaining three triads, from around 13% in *G. tomentella* T3–45% in *G. dolichocarpa* T2, with *G. tomentella* T1 intermediate at 23%.

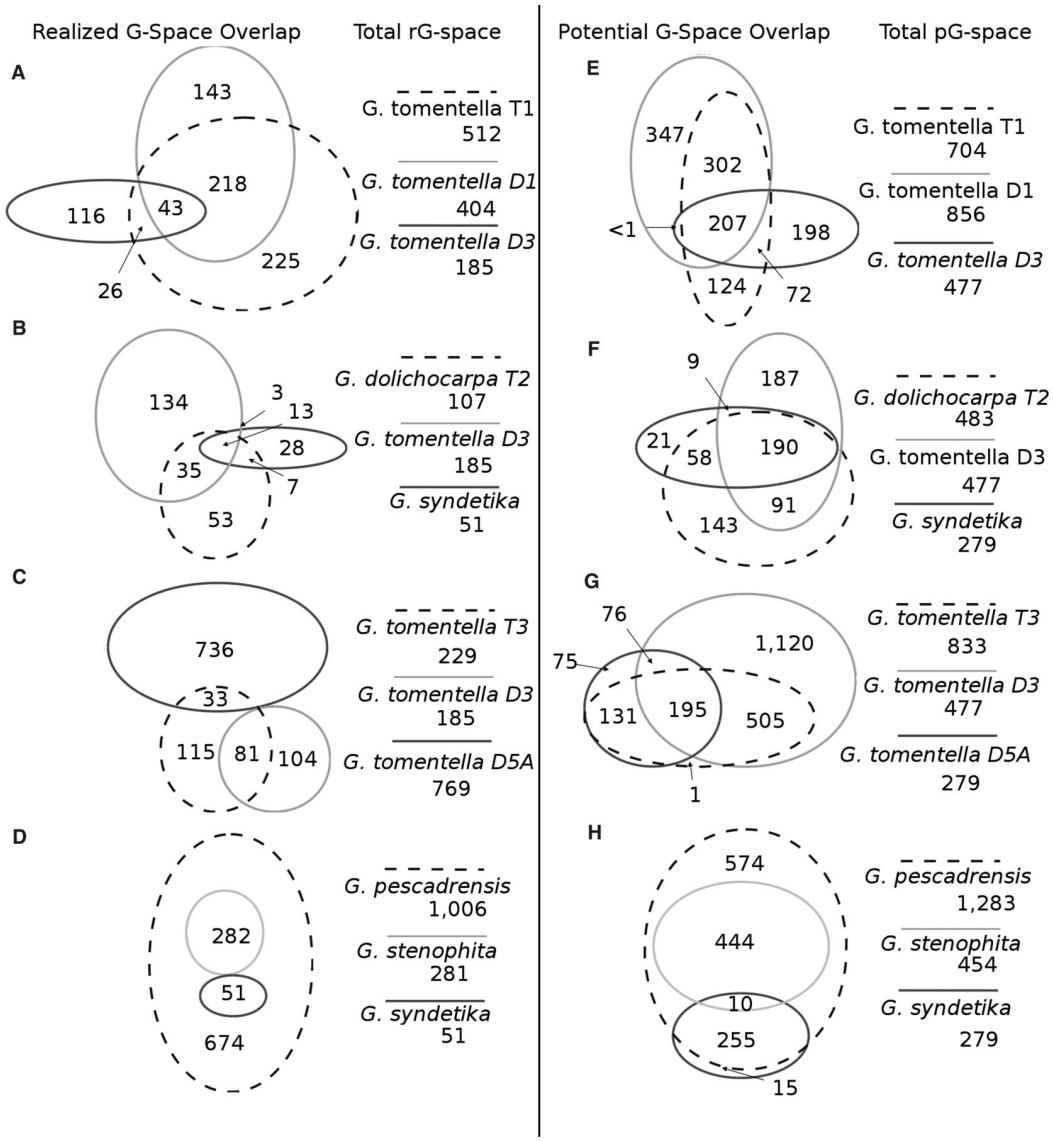

Fig. 4.3. Venn diagrams illustrating the relative niche sizes and overlaps within each polyploid triad. (A–D) Realized G-space and (E–H) potential G-space size and overlap for each polyploid triad. In each diagram, the size of the ovals corresponds to the relative geographic area represented by the realized or potential niche space for that species. All values are in thousands of square kilometers.

Whereas the previous comparisons address the question of

how much of the potential ranges of diploid progenitors are in theory accessible to the polyploid, we were also interested in knowing how similar a polyploid's pG-space is to those of its progenitors. We therefore measured the percentage contributions of each of the following components to the full range of each polyploid: (1) the area shared with both diploids (3-way sympatry); (2) the portion of polyploid pG-space shared with each diploid outside the region of 3-way sympatry; and (3) the unique polyploid pG-space, lying outside the pG-spaces of either progenitor. As in other measurements, *G. pescadrensis* was the outlier, with the zone of potential 3-way sympatry comprising only a tiny fraction of its total pG-space; in the other allopolyploids this zone comprised 23–39% of their total ranges (Fig. 4.3). Perhaps not surprisingly, the percentage of polyploid pG-spaces shared with diploid progenitors tracked the relative size of the diploid's range in all four triads, suggesting that this may not be a meaningful number. Three of the four allopolyploids (all but *G. tomentella* T3) had a substantial portion of their pG-space outside of the potential ranges of either of their diploid progenitors, with a high of 45% in the case of *G. pescadrensis* (Fig. 4.3).

Table 4.2: Schoener's D index of niche dissimilarity applied to pG-space within triads.

Comparison:	Niche similarity (D) based on: Raw Maxent score	LPT[1]-based pG-space
G. tomentella D3 vs. *G. tomentella* D1	0.31	0.23
G. tomentella D3 vs. *G. tomentella* T1*	0.45	0.38
G. tomentella D1 vs. *G. tomentella* T1*	0.63	0.59
G. tomentella D3 vs. *G. syndetika* D4	0.36	0.42
G. tomentella D3 vs. *G. dolichocarpa* T2*	0.41	0.57
G. syndetika D4 vs. *G. dolichocarpa* T2*	0.62	0.51
G. tomentella D3 vs. *G. tomentella* D5A	0.20	0.14
G. tomentella D3 vs. *G. tomentella* T3*	0.53	0.39
G. tomentella D5A vs. *G. tomentella* T3*	0.53	0.37
G. syndetika D4 vs. *G. stenophita*	0.35	0.50
G. syndetika D4 vs. *G. pescadrensis**	0.53	0.66
G. stenophita vs. *G. pescadrensis**	0.68	0.36

Note: When comparing two models, the D value ranges between "0" (representing no overlap) and "1" (representing complete overlap).) The null distribution of D values for raw Maxent scores indicates that values less than 0.84 reflect significant niche differentiation. The null distribution of D values for the LPT-based G-space indicates that values less than 0.61 reflect significant niche differentiation.

[1]- LPT – Lowest Presence Threshold

* Indicates polyploid species.

Efficiency of potential range occupancy— We define "efficiency" as the percentage of theoretically available range actually occupied by a species (rG-space/pG-space × 100; similar to

the “overlap index” of Oberprieler et al., 2012). We calculated this metric for the overall range of each species (Table 4.3, column 1), as well as in the same six biologically relevant portions of pG-space described for polyploids in the preceding section (Table 4.3, columns 2–4 and 6–8). Where relevant, we also calculated the diploid percentage occupancy of these other range components.

There was a dichotomy among the four triads for overall efficiency (Table 4.3, column 1) with high values for *G. pescadrensis* (78%) and *G. tomentella* T1 (73%) and low values for *G. dolichocarpa* T2 (22%) and *G. tomentella* T3 (28%). Efficiency in *G. pescadrensis* and *G. tomentella* T1 was considerably higher than that of either of their diploid progenitors (average 34%, range 16–60%), whereas in the other two triads one (T2 triad) or both (T3 triad) diploids had higher efficiencies than the polyploid. This wide variation in relative efficiency seen among triads was reflected in the fact that polyploids, as a group, did not have higher efficiencies than diploids (Wilcoxon signed rank = 7, P = 0.7302).

In all triads, the polyploid had a higher efficiency than either of its progenitors in the zone of 3-way sympatry (Table 4.3, column 2), but there were clear differences among the polyploids. *Glycine pescadrensis* actually occupied all of the small zone of potential 3-way sympatry and was much more efficient than either its *G. stenophita* parent, which actually occupied only 43% of this zone, or

its *G. syndetika* D4 parent (2.2%). In contrast, the other three polyploids had much lower efficiencies (41–47%) in their zones of potential progenitor sympatry, often comparable to efficiencies of the diploids (Table 4.3, column 2).

In the zones of pG-space overlap with each parent individually, *G. pescadrensis* was again far more efficient than either of its progenitors, actually occupying 87–99% of each parental pG-space as opposed to 19–62% for its diploid progenitors (Table 4.3, columns 3, 4). Though less efficient than *G. pescadrensis*, *G. tomentella* T1 (68–85%) was like it in having considerably higher efficiency than either of its progenitors in areas of potential sympatry with them. In contrast, in only one comparison was either *G. dolichocarpa* T2 or *G. tomentella* T3 even slightly more efficient than a progenitor (*G. tomentella* T3 vs. *G. tomentella* D5A, 27% vs. 26%); in the other three comparisons, the diploid had a 13–31% higher efficiency than the polyploid, though the highest diploid efficiency was only 55%.

Table 4.3: Efficiency (percent actual occupancy of potential range components).

Taxon	% Occupancy of: Total pG-space	3-Way sympatry	Sympatry with parent 1	Sympatry with parent 2	Combined diploid pG-space	Individual unique pG-space	Parent 1 pG-space	Parent 2 pG-space	Diploid – Diploid potential sympatry[1]
G. tomentella D1 (P1)	47	39	55			28			0
G. tomentella D3 (P2)	39	32		47		36			24
G. tomentella T1*	73	47	85	68	39	45	31	15	83
G. tomentella D3 (P1)	39	39	43			33			25
G. syndetika D4 (P2)	18	26		55		0			0
G. dolichocarpa T2*	22	41	30	34	16	13	18	30	100
G. tomentella D3	39	41	53			2.0			13
G. tomentella D5A	41	15		26		53			0
G. tomentella T3*	28	45	39	27	11	100	11	10	45
G. syndetika D4 (P1)	18	2.2	19			0			2.2
G. stenophita (P2)	62	43		62		No unique potential range			43
*G. pescadrensis**	79	100	87	99	91	59	91	99	100

Note(s): The column headings refer to the category of pG-space for which the percentage actually occupied (rG-space) by each species (row) is shown. Additionally the designations of “parent 1” or “P1” and “parent 2” or “P2” are made.
[1] -The “Diploid – Diploid potential sympatry” is the area of pG-space where suitable climate exists for both diploid species in a triad together.
* Indicates polyploid species.

In its efficiency over the summed diploid progenitor pG-spaces, *G. pescadrensis* again differed from the other polyploids in

actually occupying 91% of this zone (Table 4.3, column 5), with none of the other polyploids higher than 39% (*G. tomentella* T1). Breaking this down to consider the individual ranges of diploid progenitors, *G. pescadrensis* was the most efficient of any species across triads, but even so only occupied 59% of this zone (Table 4.3, column 6); neither of its parents had unique pG-space due to their overlap with *G. pescadrensis*. In the T3 triad, *G. tomentella* T3 had no unique pG-space, nor did its *G. tomentella* D3 parent; the *G. tomentella* D5A parent had an efficiency in its extensive unique pG-space nearly as high as that of *G. pescadrensis* in its triad. In the two triads where all three species had at least some unique pG-space, *G. tomentella* T1 was marginally most efficient in its triad, whereas in the T2 triad the *G. tomentella* D3 progenitor was much more efficient (33%) than the polyploid (12.5%).

Climate niche space comparisons— Although potential range is a closer approximation of underlying physiological traits than is actual range, a small change in physiology that allows a species to occupy a new habitat could result in a large increase in range size if that habitat is geographically extensive. This is certainly interesting in the context of the actual ecology of a species and is relevant to discussions of Australian biogeography in the case of *Glycine*. However, it is useful to further decouple physiology from

geographic range by estimating the sizes and positions of niches in environmental (E-) space. Within most triads, there were no significant differences in niche breadth among either diploids or polyploids (Table 4.4). Although in three of four triads, the polyploid's niche space was numerically largest, only two polyploids were significantly larger in E-space, and then only compared with one of their two parents (*G. tomentella* T1 > *G. tomentella* D1; *G. pescadrensis* > *G. stenophita*). In the PCoA analysis of all taxa, polyploids did not have significantly larger niche sizes than diploids as a group (Wilcoxon signed rank = 4, P = 0.19, results not shown).

The relative position of the centroids in each triad in the PCoA plot of the first two axes indicates how similar the niches of the various taxa are (Fig. 4.4). The two axes plotted for each triad visualized between 88.9% (T1 triad) and 97.1% (T2 triad) of the variation in E-space. In all four triads, niche positions of diploids were significantly different from one another (Table 4.5). In each case, the polyploid centroid was positioned between the centroids of its diploid progenitors. In the T1 and T2 triads the polyploid's centroid was positioned near the midpoint of the line connecting the diploid progenitor centroids (Fig. 4.4), whereas in the *G. pescadrensis* and T3 triads the polyploid's centroid was much more distant from one progenitor than the other. Taking into account statistical significance, two of the polyploids (*G. dolichocarpa* T2

and *G. tomentella* T3) were significantly differentiated from one parent (*G. syndetika* D4 and *G. tomentella* D1, respectively) but not from their shared *G. tomentella* D3 parent. *G. tomentella* T1 was significantly differentiated from both parents, whereas although the *G. pescadrensis* centroid was skewed toward *G. stenophita*, *G. pescadrensis* E-space was not significantly differentiated from that of either parent (Table 4.5).

Non-Australian ranges— Localities occupied by *Glycine* polyploids in Taiwan, West Timor, and Papua New Guinea were tested to determine whether their climates could potentially support the other *Glycine* species studied here. *Glycine pescadrensis* also occurs in the Ryukyu Islands chain of Japan, but climate modeling was not possible for these small islands. Results are reported as positive sites/total sites (Table 4.5). For example, the score of 7/7 for *G. tomentella* D3 in Taiwan means that the model for that species predicted that the climate at all seven sites on Taiwan from which other *Glycine* species were collected (including *G. tomentella* T4 and polyploid *G. tabacina*) would be suitable for *G. tomentella* D3. Results suggest that Taiwan possesses suitable climate for all of the species studied here except one diploid, *G. stenophita*; this includes all other diploids (*G. tomentella* D1, D3, D5A, and *G. syndetika* D4) and two polyploids

(*G. tomentella* T1 and T3) that do not occur there. In contrast, Papua New Guinea fits the climate niches of only two of the three species (*G. tomentella* D3 and *G. tomentella* T3; not *G. tomentella* T1) found there, whereas West Timor has a suitable climate niche only for *G. tomentella* T3, which occurs there, and *G. tomentella* D3, which does not.

Table 4.4: Comparison of similarity of niche breadth and centroid position from Principal Coordinate Analyses.

Taxon pair	Niche breadth: (Tukey HSD comparison p-value)	Mean niche position comparison: (post hoc Games-Howell p-value)
G. tomentella D3 – *G. syndetika* D4	0.10 vs. 0.05 (0.012')	0.00''
G. tomentella D3 – *G. dolichocarpa* T2*	0.10 vs. 0.09 (0.46)	0.09
G. syndetika D4 – *G. dolichocarpa* T2*	0.05 vs. 0.09 (0.20)	0.20
G. tomentella D3 – *G. tomentella* D1	0.11 vs. 0.08 (0.10)	0.00''
G. tomentella D1 – *G. tomentella* T1*	0.08 vs. 0.11 (0.04')	0.03'
G. tomentella D3 – *G. tomentella* T1*	0.11 vs. 0.11 (1.000)	0.007''
G. tomentella D3 – *G. tomentella* D5A	0.11 vs. 0.09 (0.14)	0.003''
G. tomentella D3 – *G. tomentella* T3*	0.11 vs. 0.11 (0.57)	0.40
G. tomentella D5A – *G. tomentella* T3*	0.09 vs. 0.11 (0.06)	0.006''
G. stenophita - *G. syndetika* D4	0.03 vs. 0.05 (0.63)	0.003''
G. stenophita – *G. pescadrensis**	0.03 vs. 0.07 (0.02')	0.33
G. syndetika D4 -- *G. pescadrensis**	0.05 vs. 0.07 (0.76)	0.06

' P < 0.05; '' P < 0.01

Note: Niche breadth is represented here by a pairwise comparison of average distance to centroid.

* On taxon name indicates the polyploid species of each taxon pair.

Discussion

Among the many questions that motivate the field of polyploidy research (e.g., Doyle et al., 2008; Soltis et al., 2010) is the degree to which polyploidy generates predictable results—whether there are molecular, physiological, developmental, or

ecological “rules” that govern the evolution of polyploids across different taxa. One of the most attractive attributes of the Glycine allopolyploid complex is that any strongly deterministic features of polyploidy should emerge, in the form of convergent phenotypes, against the background of relatively similar genomes brought together independently in various overlapping combinations. For example, *Glycine* allopolyploids display the expected correlation between increased DNA content and guard cell size (e.g., Beaulieu et al., 2009; Coate et al., 2012; J. E. Coate, Reed College; J. J. Doyle, unpublished data), and the increased photosynthetic rate per cell observed in many polyploids (Warner and Edwards, 1993; Ilut et al., 2012). Such consistent molecular and physiological differences across triads could result in convergent ecological patterns.

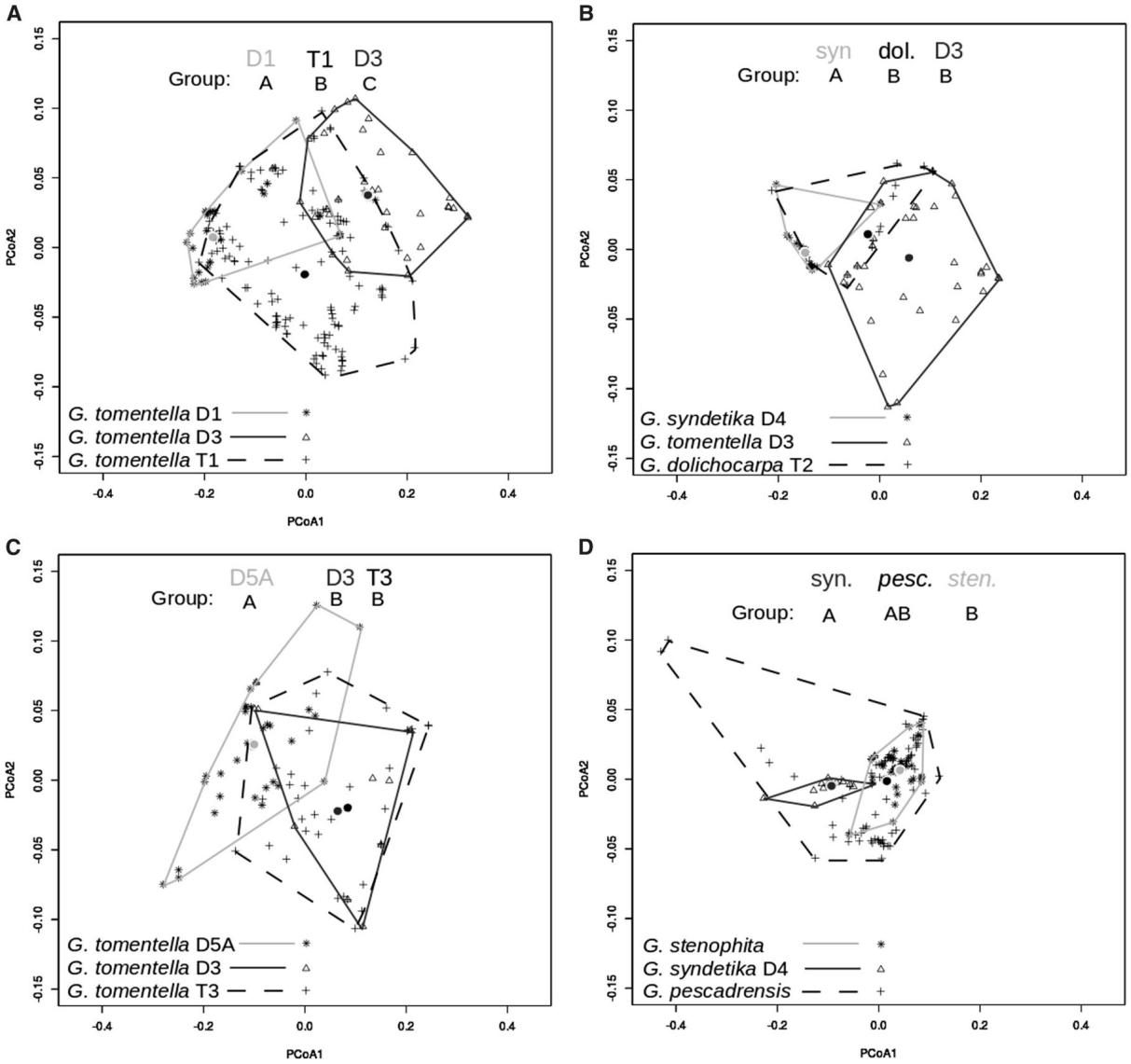

Figure 4.4. Principal coordinate analysis of occupied E-space. For *Glycine tomentella* T1 (A), *G. dolichocarpa* T2 (B), *G. tomentella* T3 (C), and *G. pescadrensis* (D) triads are shown, and the distribution of each species is outlined (see key by panel for outline and marker reference). Three dots for each panel show the geometric center (centroid) of each species' cluster and are shade matched with the distribution outlines. To aid in centroid identification, abbreviated species names are shade-coded at the top of each panel in the order the centroids appear along the x -axis. The two axes plotted for each triad visualize between 88.9% (T1 triad) and 97.1% (T2 triad) of the variation in E-space. The Games–Howell multiple comparison test was applied to centroid positions for pairwise significance testing within triads, and the significance groups are indicated by letter under the name

abbreviation at the top of each panel (for p-values, see Table 4.4).

An ecological feature strongly associated with polyploids relative to diploids is invasiveness (Pandit et al., 2011), and it also appears to be a convergent phenotype in *Glycine* allopolyploids. Whereas none of the ca. 25 diploid (2 n = 38, 40) perennial *Glycine* species are found outside of Australia and Papua New Guinea (e.g., Gonzalez-Orozco et al., 2012), five of eight known allopolyploid species have colonized islands of the Pacific Ocean (Doyle et al., 2004). This is despite the fact that the polyploids are by definition younger than their diploid progenitors and have been in existence for, at most, only a few hundred thousand years (Bombarely et al., 2014). To determine whether other convergent ecological differences exist between *Glycine* allopolyploids and their diploid progenitors, we modeled climate niches of four *Glycine* allopolyploid triads, including three with non-Australian ranges.

Table 4.5: Extra-Australian Colonization

Taxon:	Colonization of *Glycine* available sites in:		
	Taiwan	PNG	West Timor
G. tomentella D1	7/7	0/9	0/1
G. tomentella D3	7/7	8/9	1/1
G. dolichocarpa T1*	7/7	0/9	0/1
G .tomentella D3	7/7	8/9	1/1
G. syndetika D4	3/7	0/9	0/1
G. tomentella T2*	7/7[1]	0/9	0/1
G. tomentella D3	7/7	8/9	1/1
G. tomentella D5A	4/7	0/9	0/1
G. tomentella T3*	3/7	6/9[1]	1/1[1]
G. syndetika D4	3/7	0/9	0/1
G. stenophita	0/7	0/9	0/1
*G. pescadrensis**	7/7[1]	0/9	0/1

[1] - Species is know to occur on the island scored.
* Indicates polyploid species.
Format: # Suitable sites for species X/ # Total sites available to *Glycine* colonization

Glycine allopolyploids do not have consistently larger ranges or climate niches and are ecologically intermediate— As a group, *Glycine* allopolyploids resembled diploids in the sizes of their actual ranges, potential ranges, and climate niches. A similar picture emerged in comparisons within each triad. Of the four polyploids, only *G. pescadrensis* had a larger actual range and potential range than both of its progenitors, several times larger than the parent with the largest ranges, *G. stenophita*. In the other three triads, there was no consistent difference between polyploids and diploids. Although polyploids tended to have the largest climate niches in each triad, few of the differences were significant. The fact that this was true even of *G. pescadrensis* suggests that its broader geographical range is due to factors other than climate-related ecophysiology. In each triad both the potential geographical ranges and the environmental niche spaces of the two diploid progenitors were significantly differentiated from one another (Tables 2, 4), and in all four triads the allopolyploid was intermediate for both characteristics (Fig. 4.4, Table 4.2). The four allopolyploids showed less differentiation from their parents in environmental space than in potential geographical space, where all had potential ranges that

included regions not climatically available to their progenitors. These findings parallel the observation of Martin and Husband, (2009) that, in their sample of 432 species from 144 genera, range sizes of polyploids were in general more similar to those of diploids than diploids were to one another. Our finding of intermediacy is also of interest because many homoploid and allopolyploid hybrids show transgressive molecular, physiological, morphological, or ecological features (Mallet, 2007; Rieseberg and Willis, 2007). Although E-space is a very indirect measure of underlying physiology (Soberon and Nakamura, 2009; Higgins et al., 2012), our results suggest that the ecophysiology of these allopolyploids may also be intermediate, rather than transgressive.

Glycine allopolyploids differ from one another in their exploitation of potential range— A polyploid is most likely to originate in the zone of potential parental sympatry, though this need not always be true (e.g., *Gossypium hirsutum*: Cronn et al., 2002). Moreover, it is possible that this zone may not fall within the potential range of the polyploid if it was ecologically transgressive at its origin or if it has diverged substantially from its progenitors. We found that all four *Glycine* polyploids could potentially occupy at least 70% of the zone of potential sympatry between their parents, despite having pG-spaces that in all four cases are significantly

differentiated from their parents. This is consistent with their being ecologically intermediate rather than strongly transgressive.

The unique pG-space occupied by each allopolyploid could be due to unique ecological attributes resulting from initial transgressive effects (e.g., Mallet, 2007; Rieseberg and Willis, 2007) or from divergence after polyploidy (e.g., Ramsey, 2011). Of the four allopolyploids, *G. tomentella* T3 had virtually no unique potential range, whereas the other three had nonparental potential ranges that comprised varying amounts of their total potential ranges. In particular, the potential range of *G. pescadrensis* not only covered far more of its parent's pG-spaces than did the other polyploids, but its unique pG-space was very large, comprising 45% of its total potential range (Fig. 4.3).

The four polyploids also differed greatly in their efficiency in exploiting their potential ranges and those of their progenitors. Not only did *G. pescadrensis* have a very extensive pG-space, but it actually fills over three quarters of it, and the same high efficiency was observed in the components of its overall potential range—zone of potential 3-way sympatry as well as combined and individual diploid progenitor pG-spaces. This polyploid also had much higher efficiencies than its diploid progenitors in zones of sympatry with them. Interestingly, *G. pescadrensis* showed its lowest efficiency in occupation of its unique potential range; this could be due to biotic

or dispersal constraints, though its ability to disperse to Taiwan and the Ryukyu Islands suggests that it is not limited by dispersal.

Australian range and climate niche do not correlate with greater colonizing ability in Glycine allopolyploids— One reason for conducting environmental niche modeling studies on the Australian ranges of these *Glycine* allopolyploids was to determine whether characteristics of their Australian distributions could explain why it is only allopolyploids in subg. *Glycine* that have spread beyond Australia and Papua New Guinea. Of the four allopolyploid species studied here, *G. pescadrensis* stands out as having a large Australian range that encompasses those of its progenitors, and it is one of the polyploids with populations outside Australia. However, *G. dolichocarpa* T2 and *G. tomentella* T3 also have populations outside Australia and Papua New Guinea—for example, *G. dolichocarpa* T2, like *G. pescadrensis*, has colonized Taiwan—but nothing about the Australian ranges of these polyploids suggests that they should be as effective colonizers as *G. pescadrensis*. In fact, it is *G. tomentella* T1 that most closely approaches *G. pescadrensis* in characteristics of its Australian distribution, but it has no populations outside of Australia and Papua New Guinea. Our findings suggest that it is not lack of climatically suitable habitat in Taiwan that prevents any of the *Glycine* species, including diploids,

from colonizing that island. If the presence of only polyploids in Taiwan is not due to climatically related attributes, it could be due to relative dispersal abilities of the different species. However, all of these *Glycine* species have pods that shatter to disperse their small, hard seeds, and all would seem equally likely to be dispersed by the birds that are proposed agents of their dissemination outside of Australia (Hymowitz et al., 1990). *Glycine dolichocarpa* T2 is more robust and grows faster than either of its progenitors (J. J. Doyle, unpublished data), which could be advantageous in a colonist, but this would also seem advantageous in its Australian range, which is not substantially larger than the ranges of its progenitors.

Variation among Glycine allopolyploids is consistent with the "rarely successful polyploids" hypothesis— Adding to the longstanding debate about evolutionary success of polyploids (e.g., Madlung, 2013), Arrigo and Barker (2012, p. 140) recently suggested that "despite leaving a substantial legacy in plant genomes, only rare polyploids survive over the long term and most are evolutionary dead-ends." Whether this is true (see Soltis et al., in press), a necessary part of this "rarely successful polyploids" hypothesis (to quote the title of their paper) is that genome doubling, per se, does not predispose all polyploids to establish new lineages—the hallmark of ecological success as defined by Wilson

(1987). Observations that polyploids may have either larger or smaller geographical or environmental ranges than diploids (e.g., Lowry and Lester, 2006; Martin and Husband, 2009; Theodoridis et al., 2013) are consistent with this hypothesis. On the other hand, the “substantial legacy” (Arrigo and Barker, 2012) of polyploidy in all angiosperm genomes is echoed by cases such as *Solanum* (Hijmans et al., 2007) and *Leucanthemum* (Oberprieler et al., 2012), where it is polyploids that have the largest ranges, even though diploids and polyploids do not differ in aggregate. We also observe this pattern in *Glycine*. Although few generalizations can be made about the four allopolyploids studied here, one of them, *G. pescadrensis*, stands out as having a wide Australian actual range, filling its potential geographical space more efficiently than its progenitors or other allopolyploids, and having dispersed populations outside of Australia. These attributes could predispose *G. pescadrensis* for long-term evolutionary success.

Acknowledgements

The authors thank Dr. Joe Miller and Dr. Bob Godfree for comments on an early draft of the manuscript and Doug Soltis and two anonymous reviewers for helpful comments on the first submitted version. The authors also thank Dr. Kevin Nixon for his time spent discussing the various methods applied here. Several

grants from the U.S. National Science Foundation have supported the authors' research on *Glycine* over many years, currently 0822258.

REFERENCES

Ackerly, D.D. 2003. Community assembly, niche conservatism, and adaptive evolution in changing environments. *International Journal of Plant Sciences* 164(S3):S165–S184.

Adams, D.K., and A.C. Comrie. 1997. The North American Monsoon. *Bulletin of the American Meteorological Society* 78:2197–2212.

Agassiz, L. 1840. Etudes sur les glaciers. Privately published, Neuchatel.

Alley, R.B. 2000. The Younger Dryas cold interval as viewed from central Greenland. *Quaternary Science Reviews* 19:213-226.

Alley, R.B. 2004. GISP2 Ice Core Temperature and Accumulation Data. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series #2004-013. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA.

Araujo, M.B, F. Ferri-Yanez, F. Bozinovic, P.A. Marquet, F. Valladares, and S.L. Chown. 2013. Heat freezes niche evolution. *Ecology Letters* 16: 1206-1219.

Arrigo, N., and M.S. Barker. 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* 15: 140 – 146.

Bailey, I.W., and E.W. Sinnott. 1915. A botanical index of Cretaceous and Tertiary climates. *Science* 41: 831 – 834.

Bailey, I.W., and E.W. Sinnott. 1916. The climatic distribution of certain types of angiosperm leaves. *American Journal of Botany* 3: 24 – 39.

Barker, M.S., N.C. Kane, M. Matvienko, A. Kozik, W. Michelmore, S.J. Knapp and L.H. Riesberg. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445 – 2455.

Barron, J.A., L. Heusser, T. Herbert, and M. Lyle. 2003. High-resolution climatic evolution of coastal northern California during the past 16,000 years. *Paleoceanography* 18(1): 1020, doi:10.1029/2002PA000768

Beaulieu J., M. Jean, and F. Belzille. 2009. The allotetraploid *Arabidopsis thaliana – Arabidopsis lyrata* subsp. *petrea* as an alternative model system for the study of polyploidy in plants. *Molecular Genetics and Genomics* 281: 421 – 435.

Beck, J., M. Boller, A. Erhardt, and W. Schwanghart. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* 19: 10 – 15.

Bennett, M.D. 1972. Nuclear DNA content and minimum generation tune in herbaceous plants. *Proceedings of the Royal Society, B, Biological Sciences* 181: 109 – 135.

Bereiter, B., S. Eggleston, J. Schmitt, C. Nehrbass-Ahles, T.F. Stocker, H. Fischer, S. Kipfstuhl, and J. Chappelaz. 2015. Revision of the EPICA Dome C CO_2 record from 800 to 600 kyr before present. *Geophysical Research Letters*. doi: 10.1002/2014GL061957

Berger, A. L. 1978. Long-term variations of daily insolation and Quaternary climatic changes, *Journal of Atmospheric Science* 35:2362-2367.

Berger, A. 1980. The Milankovitch Astronomical Theory of Paleoclimates: A Modern Review. *Vistas in Astronomy* 24: 103-122.

Bertrand, R., J. Lenoir, C. Piedallu, G. Riofrio-Dillon, P. de Ruffray, C. Vidal, J. Pierrat, and J.C. Gégout. 2011. Changes in plant community composition lag behind climate warming in lowland forests. *Nature* 479: 517 – 520.

Betancourt, J.L. and T.R. Van Devender. 1981. Holocene Vegetation in Chaco Canyon, New Mexico. *Science* 214(4521): 656-658.

Betancourt, J.L., T.R. Van Devender, and P.S. Martin. 1990. Packrat middens: the last 40,000 years of biotic change. University of Arizona Press, Tucson, AZ.

Birchler, J., and R. Veitia. 2014. The gene balance hypothesis: Dosage effects in plants. In C. Spillane and P. C. McKeown [eds.], *Methods in molecular biology*, vol. 1112, 25–32. Humana Press. New York, New York, USA.

Blanc G., and K.H. Wolfe. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell*

16: 1679 – 1691.

Blonder, B., D. Nogués-Bravo, M. K. Borregaard, J.C. Donoghue, P.M. Jørgensen, N.J.B. Kraft, J. Lessard, et al. 2015. Linking environmental filtering and disequilibrium to biogeography with a community climate framework. *Ecology* 96: 972 – 985.

Bombarely, A., J.E. Coate, and J.J. Doyle. 2014. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. *PeerJ* 2:e391

Boyle, B.L. 1996. Changes on altitudinal and latitudinal gradients in Neotropical montane forests. Graduate School of Arts and Sciences of Washington University, St. Louis, Missouri, USA.

Brochmann, C., A. K. Brysting, I. G. Alsos, L. Borgen, H. H. Grundt, A. C. Scheen, and R. Elven. 2004. Polyploidy in arctic plants. *Biological Journal of the Linnean Society* 82: 521 – 536.

Carlson, A.E. 2013. The Younger Dryas Climate Event. *Encyclopedia of Quaternary Science* 3: 126-134.

Carozzi, A. V., Ed. 1967. Studies on Glaciers Preceded by the Discourse of Neuchdtel by Louis Agassiz. Hafner, New York.

Coate, J.E., A.K. Luciano, V. Seralathan, K.J. Minchew, T.G. Owens, and J.J. Doyle. 2012. Anatomical, biochemical, and photosynthetic responses to recent allopolyploidy in *Glycine dolichocarpa* (Fabaceae). *American Journal of Botany* 99: 55 – 67.

Cole, K.L. 2009. Vegetation Response to Early Holocene Warming as an Analog for Current and Future Changes. *Conservation Biology* 24(1), 29-37.

Cole, K.L., K. Ironside, J. Eischeid, G. Garfin, P.B. Duffy, and C. Toney. 2011. Past and ongoing shifts in Joshua tree distribution support future modeled range contraction. *Ecological Applications* 21(1): 1370149.

Cole, K.L., and S.T. Arundel. 2005. Carbon isotopes from fossil packrat pellets and elevational movements of Utah agave plants reveal the Younger Dryas cold period in Grand Canyon, Arizona. *Geology* 33(9): 713-716.

Couvreur, T. L., F. Forest, and W.J. Baker. 2011. Origin and global

diversification patterns of tropical rain forests: Inferences from a complete genus level phylogeny of palms. *BMC Biology* 9: 44.

Cronn, R.C., R.L. Small, T. Haselkorn, and J.F. Wendel. 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany* 89: 707 – 725.

Cuffey, K.M., and G.D. Clow. 1997. Temperature, accumulation, and ice sheet elevation in central Greenland through the last deglacial transition. *Journal of Geophysical Research* 102:26383-26396.

Crucifix, M. 2016. palinsol: Insolation for Palaeoclimate Studies. R package version 0.93. http://CRAN.R-project.org/package=palinsol

Daly, C., W.P. Gibson, G.H. Taylor, G.L. Johnson, and P. Pasteris. 2002. A knowledge-based approach to the statistical mapping of climate. *Climate Research* 22: 99 – 113.

Daly, C., M. Halbleib, J.I. Smith, W.P. Gibson, M.K. Doggett, G.H. Taylor, J. Curtis, and P.P. Pasteris. 2008.Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology* 28: 2031 – 2064.

Daly, C., G.H. Taylor, W.P. Gibson, T.W. Parzybok, G.L. Johnson, and P.A. Pasteris. 2000. High-quality spatial climate data sets for the United States and beyond. *Transactions of the American Society of Agricultural Engineers* 43: 1957 – 1962.

Day, R.W., and G.P. Quinn . 1989. Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs* 59 : 433 – 463 .

De Frenne, P., F. Rodriguez-Sanchez, D.A. Coomes, L. Baeten, G. Verstraeten, M. Vellend, M. Berhnhardt-Romermann, et al. 2013. Microclimate moderates plant responses to macroclimate warming. *Proceedings of the National Academy of Sciences, USA* 110: 18561 – 18565.

Delong, E.R., E.M. Delong, and D. L. Clarke-Pearson. 1988. Comparing the area under two or more correlated receiver operator characteristic curves: A non-parametric approach. *Biometrics* 44: 837 – 845.

Dial, K.P. and N.J. Czaplewski. 1990. Do woodrat middens accurately represent the animals' environments and diets? The Woodhouse Mesa study. Packrat middens: the last 40,000 years of biotic change (ed. By J.L. Betancourt, T.R. Van Devender and P.S. Martin), pp. 43–58. University of Arizona Press, Tucson, AZ.

Diaz, S., M. Cabido, and F. Casanoves. 1998. Plant functional traits and environmental filters at a regional scale. *Journal of Vegetation Science* 9: 113 – 122.

Doyle, J.J., J.L. Doyle, and A.H.D. Brown. 1999. Origins, colonization, and lineage recombination in a widespread perennial soybean polyploid complex. *Proceedings of the National Academy of Sciences*, USA 96: 10741 – 10745.

Doyle, J.J., J.L. Doyle, A.H.D. Brown, and R.G. Palmer. 2002. Genomes, multiple origins, and lineage recombination in the *Glycine tomentella* (Leguminosae) polyploid complex: Histone H3-D gene sequences. *Evolution* 56: 1388 – 1402.

Doyle, J.J., J.L. Doyle, J.T. Rauscher, and A.H.D. Brown. 2004. Evolution of the perennial soybean polyploid complex (*Glycine* subgenus *Glycine*): A study of contrasts. *Biological Journal of the Linnean Society* 82: 583 – 597.

Doyle, J.J., L.E. Flagel, A.H. Paterson, R.A. Rapp D.E. Soltis, P.S. Soltis, and J.F. Wendel. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics* 42: 443 – 461.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129 – 151.

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

Engemann, K., B. J. Enquist, B. Sandel, B. Boyle, P. M. Jorgensen, N. Morueta-Holme, C. Violle, and J. Svenning. 2015. Limited sampling hampers "big data" estimation of richness in a tropical biodiversity hotspot. *Ecology and Evolution* 5(3): 807-820. doi: 10.1002/ece3.1405

Feeley, K.J., and M.R. Silman. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity & Distributions* 17: 1132 – 1140.

Finley, R.B. 1990. Woodrat Ecology and Behavior and the Interpretation of Paleomiddens. Packrat middens: the last 40,000 years of biotic change (ed. By J.L. Betancourt, T.R. Van Devender and P.S. Martin), pp. 28-43. University of Arizona Press, Tucson, AZ.

Francis, D., M.S. Davies, and P.W. Barlow. 2008. A strong nucleotypic effect on the cell cycle regardless of ploidy level. *Annals of Botany* 101: 747 – 757.

Freeling, M., M.R. Woodhouse, S. Subramamian, G. Turco, D. Lisch, and J. C. Schnable. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* 15: 131 – 139.

Games, P.A. , and J.E. Howell. 1976. Pairwise multiple comparison procedures with unequal N ’s and/or variances: A Monte-Carlo study. *Journal of Educational Statistics* 1:113 – 125 .

Gentry, A. H. 1988. Changes in plant community diversity and floristic composition on environmental and geographic gradients. *Annals of the Missouri Botanical Garden* 75: 1 – 34.

Gleason, H.A. 1926. The individualistic concept of plant association. *Bulletin of the Torrey Botanical Club* 53: 7 – 26.

Gleason, H.A. 1939. The individualistic Concept of the Plant Association. *The American Midland Naturalist* 21(1): 92 - 110.

Godsoe W., M.A. Larson, K.L. Glennon, and K.A. Segraves. 2013. Polyploidization in *Heuchera cylindrica* (Saxifragaceae) did not result in a shift in climatic requirements. *American Journal of Botany* 100: 496 – 508.

Gonzalez-Orozco, C.E., A.H.D. Brown, N. Knerr, J.T. Miller, and J.J. Doyle. 2012. Hotspots of diversity of wild Australian soybean relatives and their conservation in situ. *Conservation Genetics* 13: 1269 – 1281.

Greenwood, D.R., P. Wilf, S.L. Wing, and D.C. Christophel. 2004. Paleotemperature estimation using leaf-margin analysis: Is

Australia different? *Palaios* 19: 129 – 142.

Grimm, G.W., and T. Denk. 2012. Reliability and resolution of the Coexistence Approach—a revalidation using modern-day data. *Review of Palaeobotany and Palynology* 172: 33 – 47.

Grimm, G.W., and A.J. Potts. 2015. Fallacies and fantasies: the theoretical underpinnings of the Coexistence Approach for palaeoclimate reconstruction. *Climate of the Past Discussions* 11: 5727-5754.

Guralnick, R.P., J. Wieczorek, R. Beaman, R.J. Hijmans, and The BioGeomancer Working Group. 2006. BioGeomancer: Automated georeferencing to map the world's biodiversity data. *PLoS Biology* 4: e381.

Guthrie, R.D. 2006. New carbon dates link climatic change with human colonization and Pleistocene extinctions. *Nature* 441: 207-209.

Hadly E.A., P.A. Spaeth, and C. Li. 2009. Niche conservatism above the species level. *Proceedings of the National Academy, USA* 106:19707–19714.

Hayter, A.J. 1984. A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative. Annals of Statistics 12: 61 – 75.

Hernandez, P.A., C.H. Graham, L.L. Master, and D.L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29: 773 – 785.

Herman, A.B., and R.A. Spicer. 1996. Palaeobotanical evidence for a warm Cretaceous Arctic Ocean. *Nature* 380: 330 – 333.

Herman, A.B., and R.A. Spicer. 1997. New quantitative palaeoclimate data for the Late Cretaceous Arctic: Evidence for a warm polar ocean. *Palaeogeography, Palaeoclimatology, Palaeoecology* 128: 227 – 251.

Higgins, S.I., R.B. O'Hara, O. Bykova, M.D. Cramer, I. Chuine, E. Gerstner, T. Hickler, et al. 2012. A physiological analogy of the niche for projecting the potential distribution of plants. *Journal of Biogeography* 39: 2132 – 2145.

Hijmans, R.J. 2014. Raster: Geographic data analysis and modeling. R package version 2.2-31. http://CRAN.R-project.org/package=raster

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

Hijmans, R.J., T. Gavrilenko, S. Stephenson, J. Bamberg, A. Salas, and D.M. Spooner. 2007. Geographical and environmental range expansion through polyploidy in wild potatoes (*Solanum* section Petota). *Global Ecology and Biogeography* 16: 485 – 495.

Hijmans, R., S. Phillips, J. Leathwick, and J. Elith. 2012. dismo: Species distribution modeling. R package version 0.7-23. http://CRAN.R-project.org/package=dismo.

Hill, A.W., R. Guralnick, P. Flemons, R. Beaman, J. Wieczorek, A. Ranipeta, V. Chavan, and D. Remsen. 2009. Location, location, location: Utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinformatics* 10 (Supplement 14): S3.

Holdridge, L.R. 1947. Determination of world plant formations from simple climatic data. *Science* 105: 367 – 368.

Holmgren, C.A., J.L. Betancourt, M.C. Penalba, J. Delgadillo, K. Zuravnsky, K.L. Hunter, K.A. Rylander, and J.L. Weiss. 2014. Evidence against a Pleistocene desert refugium in the Lower Colorado River Basin. *Journal of Biogeography* 41(9): 1769-1780.

Huntley, B., P.J. Bartlein, and I.C. Prentice. 1989. Climatic control of the distribution and abundance of beech (*Fagus* L.) in Europe and North America. *Journal of Biogeography* 16:551–560.

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology 22: 415 – 427.

Hymowitz T., R.J. Singh, and R.P. Larkin. 1990. Long-distance dispersal: The case for the allopolyploid *Glycine tabacina* (Labill.) Benth. and *Glycine tomentella* Hayata in the west-central Pacific. *Micronesica* 23: 5 – 14.

Ibarra, D.E., A.E. Egger, K.L. Weaver, C.R. Harris, and K. Maher.

2014. Rise and fall of late Pleistocene pluvial lakes in response to reduced evaporation and precipitation: Evidence from Lake Surprise, California. *Geological Society of America Bulletin* 11(12), 1387–1415.

Ilut, D.C., J.E. Coate, A.K. Luciano, T.G. Owens, G.D. May, A. Farmer, and J.J. Doyle. 2012. A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. *American Journal of Botany* 99: 383 – 396.

Imbrie, J., J. Imbrie, E.A. Boyle, S.C. Clemens, A. Duffy, W.R. Howard, G. Kukla, J. Kutzbach, D.G. Martinson, A. Mcintyre, A.C. Mix, B. Molfino, J.J. Morley, L.C. Peterson, N.G. Pisias, W.L. Prell, M.E. Raytoo, N.J. Shackleton, and J.R. Toggweiler. 1992. On the structure and origin of major glaciation cycles: I. Linear responses to Milankovitch forcing. *Paleoceanography* 7:701–738.

Jacques, F.M.B., T. Su, R.A. Spicer, Y. Xing, Y. Huang, W. Wang, and Z. Zhou. 2011. Leaf physiognomy and climate: Are monsoon systems different? *Global and Planetary Change* 76: 56 – 62.

Jenkins, J., and G. Motzkin. 2009. Harvard Forest flora database from 1980 to present. Harvard Forest Data Archive: HF116, http://harvardforest.fas. harvard.edu/data/p11/hf116/hf116-01-flora.csv.

Jiao, Y., N.J. Wickett, S. Ayyampalayam, A.S. Chanderbali, L. Landherr, P. E. Ralph, L.P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97 – 100.

Kennedy, E.M., R.A. Spicer, and P.M. Rees. 2002. Quantitative paleoclimate estimates from Late Cretaceous and Paleocene leaf floras in the northwest of the South Island, New Zealand. *Palaeogeography, Palaeoclimatology, Palaeoecology* 184: 321 – 345.

Killeen, T.J., and T.S. Schulenberg. 1998. A biological assessment of Parque Nacional Noel Kempff Mercado, Bolivia. *RAP Working Papers*, vol. 10, pp. 61–85. Conservation International, Washington, D.C., USA.

Kowalski, E.A., and D.L. Dilcher. 2003. Warmer paleotemperatures for terrestrial ecosystems. *Proceedings of the National Academy of Sciences*, USA 100: 167 – 170.

Kühl, N., C. Gebhardt, T. Litt, and A. Hense. 2002. Probability density functions as botanical-climatological transfer functions for climate reconstruction. *Quaternary Research* 58: 381 – 392.

Lachniet, M.S., R.F. Denniston, Y. Asmerom, and V.J. Polyak. 2014. Orbital control of western North America atmospheric circulation and climate over two glacial cycles. *Nature Communications* 5: 3805

LaMarche , V.C. 1973 . Holocene climatic variations inferred from tree-line fluctuations in the White Mountains, California. *Quaternary Research* 3 : 632 – 660 .

Leitch, I.J., and M.D. Bennett. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82: 651 – 663.

Lenoir , J. , J.C. Gégout , P.A. Marquet , P. de Ruff ray , and H. Brisse . 2008 . A significant upward shift in plant species optimum elevation during the 20th century. *Science* 320: 1768 – 1771.

Little, S.A., S.W. Kembel, and P. Wilf. 2010. Paleotemperature proxies from leaf fossils reinterpreted in light of evolutionary history. *PLoS ONE* 5: e15161.

Liu, C., M. White, and G. Newell. 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography* 34: 232 – 243.

Losos, J.B. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters* 11: 995 – 1003.

Lowry, E., and S.E. Lester. 2006. The biogeography of plant reproduction: Potential determinants of species' range sizes. *Journal of Biogeography* 33: 1975 – 1982.

MacDonald, G.M., K.A., Moser, A.M. Bloom, D.F. Porinchu, A.P. Potito, B.B. Wolfe, T.W.D. Edwards, A. Pateel, A.R. Orme, and A.J. Orme. 2008. Evidence of temperature depression and hydrological variations in the eastern Sierra Nevada during the Younger Dryas stade. *Quaternary Research* 70: 131-140

Madlung, A. 2013. Polyploidy and its effect on evolutionary success: Old questions revisited with new tools. *Heredity* 110: 99 – 104.

Mallett, J. 2007. Hybrid speciation. *Nature* 446: 279 – 283.

Mann, M.E., Z. Zhang, S. Rutherford, R.S. Bradley, M.K. Hughes, D. Shindell, C. Ammann, G. Faluvegi, and F. Ni. Global Signatures and Dynamical Origins of the Little Ice Age and Medieval Climate Anomaly. *Science* 326: 1256-1260.

Martínez-Meyer, E., and A.T. Peterson. 2006. Conservatism of ecological niche characteristics in North American plant species over the Pleistocene-to-Recent transition. *Journal of Biogeography* 33: 1779 – 1789.

Martin, P.S. and C.M. Drew. 1969. Scanning Electron Photomicrographs of Southwestern Pollen Grains. *Journal of the Arizona Academy of Sciences* 5(3): 147-176.

Martin, S.L., and B.C. Husband. 2009. Influence of phylogeny and ploidy on species ranges of North American angiosperms. *Journal of Ecology* 97: 913 – 922.

Mateo, R.G., T.B. Croat, A.M. Felicismo, and J. Munoz. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absence and target-group absences from natural history collections. *Diversity & Distributions* 16: 84 – 94.

McIntyer, P.J. 2012. Polyploidy associated with altered and broader ecological niches in the *Claytonia perfoliata* (Portulacaceae) species complex. *American Journal of Botany* 99: 655 – 662.

Meimberg, H., K.J. Rice, N. F. Milan, C.C. Njoku, and J.K. McKay. 2009. Multiple origins promote the ecological amplitude of allopolyploid *Aegilops* (Poaceae). *American Journal of Botany* 96: 1262 – 1273.

Meyers, L.A., and D.A. Levin. 2006. On the abundance of polyploids in flowering plants. *Evolution* 60: 1198 – 1206.

Mosbrugger, V., and T. Utescher. 1997. The Coexistence Approach—a method for quantitative reconstructions of Tertiary terrestrial palaeoclimate data using plant fossils. *Palaeogeography, Palaeoclimatology, Palaeoecology* 134: 61 – 86.

New, M., D. Lister, M. Hulme, and I. Makin. 2002. A high-resolution

data set of surface climate over global land areas. *Climate Research* 21: 1 – 25.

Oberprieler, C., K. Konowalik, S. Altpeter, E. Siegert, R.M. Lopresti, R. Greiner, and R. Vogt. 2012. Filling of eco-climatological niches in a polyploid complex—A case study in the plant genus *Leucanthemum* Mill. (Compositae, Anthemideae) from the Iberian Peninsula. *Flora (Jena)* 207: 862 – 867.

Oksagen, J., F.G. Blanchet, R. Kindt, P. Legendre, P.R. Minchin, R.B. O'Hara, G.L. Simpson, et al. 2012. vegan: Community ecology package. R package version 2.0-5. http://CRAN.R-project.org/package=vegan.

Oster, J.L., D.E. Eibarra, M.J. Winnick, and K. Maher. 2015. Steering of westerly storms over western North America at the Last Glacial Maximum. *Nature Geoscience* 8, 201-205.

Otto, S.P., and J. Whitton. 2000. Polyploidy incidence and evolution. *Annual Review of Genetics* 34: 401 – 437.

Paciorek, C.J., and J.S. McLachlan. 2009. Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *Journal of the American Statistical Association* 104: 608 – 622.

Pandit M.K., M.J.O. Pocock, and W.E. Kunin. 2011. Ploidy influences rarity and invasiveness in plants. *Journal of Ecology* 99: 1108 – 1115.

Parnell, A. 2015. Bchron: Radiocarbon Dating, Age-Depth Modelling, Relative Sea Level Rate Estimation, and Non-Parametric Phase Modelling. R package version 4.1.2. http://CRAN.R-project.org/package=Bchron

Paterson, A.H., B.A. Chapman, J.C. Kissinger, J.E. Bowers, F.A. Feltus, and J.C. Estill. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends in Genetics* 22: 597 – 602.

Pearman, P.B., A. Guisan, O. Broennimann, and C.F. Randin. 2008. Niche dynamics in space and time. *Trends in Ecology & Evolution* 23: 149 – 158.

Pearse, I.S., T. Kruegel, and I.T. Baldwin. 2006. Innovation in anti-herbivore defense systems during neopolypoloidy—The functional consequences of instantaneous speciation. *Plant Journal* 47: 196 – 210.

Pearson, R.G., C.J. Raxworthy, M. Nakamura, and A.T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34: 102 – 117.

Peppe, D.J., D.L. Royer, B. Cariglino, S.Y. Oliver, S. Newman, E. Leight, G. Enikolopov, et al. 2011. Sensitivity of leaf size and shape to climate: Global patterns and paleoclimatic applications. *New Phytologist* 190: 724 – 739.

Peterson, A.T., J. Soberon, R.G. Pearson, E. Martínez-Meyer, M. Nakamura, and M.B. Araujo. 2011. Ecological niches and geographic distributions. Princeton University Press, Princeton, New Jersey, USA.

Petit, C., and J.D. Thompson. 1999. Species diversity and ecological range in relation to ploidy level in the flora of the Pyrenees. *Evolutionary Ecology* 13: 45 – 66.

Phillips, O.L., and J.S. Miller. 2002. Global patterns of plant diversity: Alwyn H. Gentry's Forest Transect Data Set. Missouri Botanical Garden Press, St. Louis, Missouri, USA.

Phillips, S.J., M. Dudik, and R.E. Schapire. 2004. A maximum entropy approach to species distribution modeling. In *Proceedings of the 21st International Conference on Machine Learning*, 6550-662. Banff, Alberta, Canada.

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

Phillips, S.J., and M. Dudik. 2008. Modeling of species distributions with
Maxent: New extensions and a comprehensive evaluation. *Ecography* 31: 161 – 175.

Phillips, S.J., M. Dudik, J. Elith, C.H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: Implications for background and

pseudo-absence data. *Ecological Applications* 19: 181 – 197.

Poore, R.Z., M.J. Pavich, and H.D. Grissino-Mayer. 2005. Record of the North American southwest monsoon from Gulf of Mexico sediment cores. *Geology* 33:209–212.

Prentice, C., P.J. Bartlein, and T. Webb. 1991. Vegetation and climate change in eastern North America since the Last Glacial Maximum. Ecology 72: 2038 – 2056.

Punyasena, S.W. 2008. Estimating Neotropical paleotemperature and paleoprecipitation using plant family climate optima. *Palaeogeography, Palaeoclimatology, Palaeoecology* 265: 226 – 237.

R Development Core Team. 2011. R: A language and environment for statistical computing. Version R 2.14.1. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsey, J. 2011. Polyploidy and ecological adaptation in wild yarrow. *Proceedings of the National Academy of Sciences*, USA 108: 7096 – 7101.

Reheis, M. 1999. Extent of Pleistocene Lakes in the Western Great Basin. *USGS Miscellaneous Field Studies Map* MF-2323, Denver, CO.

Reimer, P. et al. 2013. IntCal13 and Marine13 radiocarbon age calibration curves 0-50,000 years cal BP. *Radiocarbon* 55: 1869-1887.

Rieseberg, L.H., and J.H. Willis. 2007. Plant speciation. *Science* 317: 910 – 914.

Roth-Nebelsick, A., T. Utescher, V. Mosbrugger, L. Diester-Haass, and H. Walther. 2004. Changes in atmospheric CO2 concentrations and climate from the Late Eocene to Early Miocene: Paleobotanical reconstruction based on fossil floras from Saxony, Germany.

Palaeogeography, Palaeoclimatology, Palaeoecology 205: 43 – 67.

Royer, D.L., P. Wilf, D.A. Janesko, E.A. Kowalski, and D.L. Dilcher. 2005. Correlations of climate and plant ecology to leaf size and shape: Potential proxies for the fossil record. *American Journal of Botany* 92: 1141 – 1151.

Ruddiman, W.F., D.Q. Fuller, J.E. Kutzbach, P.C. Tzedakis, J.O. Kaplan, E.C. Ellis, S.J. Vavrus, C.N. Roberts, R. Fyfe, F. He, C. Lemmen, and J. Woodridge. 2016. Late Holocene climate: Natural or Anthropogenic. *Review of Geophysics* 54, doi:10.1002/2015RG000503.

Schoener, T.W. 1968. The Anolis lizards of Bimini: Resource partitioning in a complex fauna. *Ecology* 49: 704 – 726.

Scuderi, L.A. 1987. Late-Holocene upper timberline variation in the southern Sierra Nevada. *Nature* 325: 242 – 244.

Sexton, J.P., P.J. McIntyre, A.L. Angert, and K.J. Rice. 2009. Evolution of Ecology of Species Range Limits. *Annual Review of Ecology, Evolution, and Systematics* 40: 415-436.

Sharpe, S.E. 2002. Constructing seasonal climograph overlap envelopes from Holocene packrat midden contents, Dinosaur National Monument, Colorado. *Quaternary Research* 57: 306 – 313.

Silverman, B.W. 1986. Density estimation for statistics and data analysis. Chapman and Hall, London, UK.

Sinka, K.J., and T.C. Atkinson. 1999. A mutual climatic range method for reconstructing paleoclimate from plant remains. *Journal of the Geological Society* 156: 381 – 396.

Smith, S.A., and J.M. Beaulieu. 2009. Life history influences rates of climatic niche evolution in flowering plants. *Proceedings of the Royal Society of London, Series B* 276: 4345 – 4352.

Soberon, J., and M. Nakamura. 2009. Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, USA 106: 19644 – 19650.

Soltis, D.E., V.A. Albert, J. Leebens-Mack, C.D. Bell, A.H. Paterson, C. Zheng, D. Sankoff, et al. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336 – 348.

Soltis, D.E., R.J.A. BUGGS, J.J. Doyle, and P.S. Soltis. 2010. What we still don’t know about polyploidy. *Taxon* 59: 1387 – 1403.

Soltis, D.E., C.M. Segovia-Salcedo, I. Jordon-Thaden, L.C. Majure, N.M. Miles, E. V. Mavrodiev, W. Mei, et al. In press. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. 2011. *New Phytologist.*

Spicer, R.A., P.J. Valdes, T.E.V. Spicer, H.J. Craggs, G. Strivastava, R.C. Mehrotra, and J. Yang. 2009. New developments in CLAMP: Calibration using global gridded meteorological data. *Palaeogeography, Palaeoclimatology, Palaeoecology* 283: 91 – 98.

Srivastava, G., R.A. Spicer, T.E.V. Spicer, J. Yang, M. Kumar, R. Mehrotra, and N. Mehrotra. 2012. Megaflora and palaeoclimate of a Late Oligocene tropical delta, Makum Coalfield, Assam: Evidence for the early development of the South Asia Monsoon. *Palaeogeography, Palaeoclimatology, Palaeoecology* 342–343: 130 – 142.

Stebbins, G.L., and J.C. Dawe. 1987. Polyploidy and distribution in the European flora: A reappraisal. *Botanische Jahrbucher fur Systematik, Pflanzengeschichte und Pflanzengeographie* 108: 343 – 354.

Syfert, M.M., M.J. Smith, and D.A. Coomes. 2013. The effects of sampling bias and model complexity on the performance of MaxEnt species distribution models. *PLoS ONE* 8(2): e55158. doi:10.1371/journal.pone.0055158

Te Beest, M., J.J. Roux, D.M. Richardson, A.K. Brysting, J. Suda, M. Kubešová, and P. Pyšek. 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany* 109: 19 – 45.

Theodoridis, S., C. Randin, O. Broennimann, T. Patsiou, and E. Conti. 2013. Divergent and narrower climatic niches characterize polyploid species of European primroses in *Primula* sect. Aleuritia. *Journal of Biogeography* 40: 1278–1289.

Thompson, R.S., C. Whitlock, P.J. Bartlein, S.P. Harrison, and W.G. Spaulding. 1993. Climatic changes in the western United States since 18,000 yr B.P., in Wright H.E., Kutzbach, J.E., Webb, T., III, Ruddiman,W.F., Street-Perrott, F.A., and Bartlein, P.J., eds., Global

Climates Since the Last Glacial Maximum: Minneapolis, University of Minnesota Press, p. 468–513.

Thompson, R.S., K.H. Anderson, and P.J. Bartlein. 2008. Quantitative estimation of bioclimatic parameters from presence/absence vegetation data in North America by the modern analog technique. *Quaternary Science Reviews* 27: 1234 – 1254.

Thompson, R.S., K.H. Anderson, R.T. Pelltier, L.E. Strickland, P.J. Bartlein, and S.L. Shafer. 2012. Quantitative estimation of climatic parameters from vegetation data in North America by the mutual climatic range technique. *Quaternary Science Reviews* 51: 18 – 39.

Thornton, P.E., S.W. Running, and M.A. White. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology (Amsterdam)* 190: 214 – 251.

Thornthwaite, C.W. 1948. An approach toward a rational classification of climate. *Geographical Review* 38 (1): 55–94.

Tiffney, B.H. 2008. Phylogeography, fossils, and northern hemisphere biogeography: The role of physiological uniformitarianism, *Annals of the Missouri Botanical Garden* 95:135–143.

Tiffney, B.H. and S.R. Manchester. 2001. The use of geological and paleontological evidence in evaluating plant phylogeographic hypotheses in the Northern Hemisphere Tertiary. *International Journal of Plant Science* 162:S3–S17.

Tukey, J.W. 1953. The problem of multiple comparisons. In The collected works of John W. Tukey VIII. Multiple comparisons: 1948–1983. Chapman and Hall, New York, New York, USA.

Utescher, T., A.A. Bruch, B. Erdei, I. François, D. Ivanov, F.M.B. Jacques, A.K. Kern, Y. Liu, V. Mosbrugger, and R.A. Spicer. 2014. The Coexistence Approach – Theoretical back- ground and practical considerations of using plant fossils for climate quantification. *Palaeogeography and. Palaeoclimatology* 410, 58–73.

Utescher, T., and V. Mosbrugger. 2015. The Palaeoflora Database. at http://www.geologie.unibonn.de/Palaeoflora

Van Devender, T.R., and W.G. Spaulding.1979. Development of Vegetation and Climate in the Southwestern United States. *Science*

204(4394): 701-710.

Van Devender, T.R. 1977. Holocene Woodlands in the Southwestern Deserts. *Science* 198(4313): 189-192.

Velasco-de León, M. P., R. A. Spicer, and D. C. Steart. 2010. Climatic reconstruction of two Pliocene floras from Mexico. *Palaeobiodiversity and Palaeoenvironments* 90: 99 – 110.

Veloz, S.D., J.W. Williams, J.L. Blois, F. He, B. Otto-Bliesner, and Z. Liu. 2012. No-analog climates and shifting realized niches during the late quaternary: Implications for 21st-century predictions by species distribution models. *Global Change Biology* 18: 1698 – 1713.

Wake, D.B., E.A. Hadly, and D.D. Ackerly. 2009. Biogeography, changing climates, and niche evolution: Biogeography, changing climates, and niche evolution. *Proceedings of the National Academy of Science, USA* 106(Suppl 2):19631–19636

Walter, H. 1973. Vegetation of the Earth in relation to climate and the ecophysiological conditions. Springer-Verlag, New York, New York, USA.

Warner, D.A., and G.E. Edwards. 1993. Effects of polyploidy on photosynthesis. *Photosynthesis Research* 35: 135 – 147.

Warren, D.L. 2012. In defense of 'niche modeling'. *Trends in Ecology & Evolution* 27: 497 – 500.

Warren, D.L., R.E. Glor, and M. Turelli. 2008. Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution* 62: 2868 – 2883.

Webb, R.H., M.B. Murov, T. C. Esque, D. E. Boyer, L. A. DeFalco, D. F. Haines, D. Oldershaw, et al. 2003. Perennial vegetation data from permanent plots on the Nevada Test Site, Nye County, Nevada. Open-File Report 03-336. U.S.Geological Survey, Tuscon, Arizona, USA.

Wells, P.V., and R. Berger. 1967. Late Pleistocene History of Coniferous Woodland in the Mohave Desert. *Science* 155(3770): 1640-1647.

Wells, P.V., and C.D. Jorgensen. 1964. Pleistocene Wood Rat

Middens and Climatic Change in the Mohave Desert: A Record of Juniper Woodlands. *Science* 143(3611): 1171-1173.

Whittaker, R.H. 1956. Vegetation of the Great Smoky Mountains. *Ecological Monographs* 26: 1 – 80.

Whittaker, R.H. 1967. Gradient analysis of vegetation. *Biological Reviews of the Cambridge Philosophical Society* 42: 207 – 264.

Wilf, P. 1997. When are leaves good thermometers? A new case for leaf margin analysis. *Paleobiology* 23: 373 – 390.

Wilson, E.O. 1987. Causes of ecological success: The case of the ants. *Journal of Animal Ecology* 56: 1 – 9.

Wolfe, J.A. 1993. A method of obtaining climatic parameters from leaf assemblages. U.S. Geological Survey Bulletin 2040: 1 – 71.

Wolfe, J.A. 1995. Paleoclimatic estimates from Tertiary leaf assemblages. *Annual Review of Earth and Planetary Sciences* 23: 119 – 142.

Wood, T.E., N. Takebayashi, M.S. Barker, I. Mayrose, P.B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* 106: 13875 – 13879.

Wright, W.E., A. Long, A.C. Comrie, S.W. Leavitt, T. Cavazos, and C. Eastoe. 2001. Monsoonal moisture sources revealed using temperature, precipitation, and precipitation stable isotope time series. *Geophysical Research Letters* 28:787–790.

Wurster, C.M., W.P. Patterson, D.A. McFarlane, L.I. Wassenaar, K.A. Hobson, N.B. Athhfield, and M.I. Bird. 2008. Stable carbon and hydrogen isotopes from bat guano in the Grand Canyon, USA, reveal Younger Dryas and 8.2ka events. *Geology* 36(9): 683-686.

Yang, J., Y. Wang, R. A. Spicer, V. Mosbrugger, C. Li, and Q. Sun. 2007. Climatic reconstruction at the Miocene Shanwang basin, China, using leaf margin analysis, CLAMP, coexistence approach, and overlapping distribution analysis. *American Journal of Botany* 94: 599 – 608.

Yesson, C., P.W. Brewer, T. Sutton, N. Caithness, J.S. Pahwa, M. Burgess, W. A. Gray, et al. 2007. How global is the Global

Biodiversity Informatics Facility? *PLoS One* 11: e1124. doi:10.1371/journal.pone.0001124

APPENDIX1

Table S1.1 –

http://www.amjbot.org/content/suppl/2015/08/11/ajb.1400500.DC1/Harbert_AppS1.txt

Table S1.2 –

http://www.amjbot.org/content/suppl/2015/08/11/ajb.1400500.DC1/Harbert_AppS2.doc

Table S1.3 –

http://www.amjbot.org/content/suppl/2015/08/11/ajb.1400500.DC1/Harbert_AppS3.doc

Table S1.4 –

http://www.amjbot.org/content/suppl/2015/08/11/ajb.1400500.DC1/Harbert_AppS4.xls

```
#Appendix S2.1. Second generation R-Script for CRACLE.
#########################################################
#####Climate Reconstruction Analysis using ###########
######## Coexistence Likelihood Estimation v2.0 #######
################### JULY, 2016############################
###############Robert S. Harbert#########################
###NOTES:
##This script works on Linux/Ubuntu 14.04LTS running R v3.0.2 (MAY NOT WORK WITH
OTHER VERSIONS OF R)
##Requires the R library "dismo", "plyr", and "rdgal". Each of these may be installed by
typing (uncommented):
#install.packages('dismo')
#install.packages('plyr')
#install.packages('rgdal')
##A download of GBIF records for all taxa at your site should be placed in a subfolder of
your working directory
##GBIF tab delimited files may have problems being read into R this way. Sometimes it is
necessary to pre-process these files with perl or your favorite programming language to
trim columns from the original file.
##Development of this code as an R library is underway at the time of publication. Look
for these tools on CRAN later.##
##RUN this to read in your occurrence data set:
occurrence <- read.table("gbif/occurrence.txt", head=T, sep="\t");
tax <- cbind(as.character(occurrence$scientific_name), occurrence$id, occurrence$lat,
occurrence$lon);
colnames(tax) <- c('tax', 'ind_id', 'lat', 'lon')
tax <- as.data.frame(tax)
###
##TO ENTER SITE LOCATION::
site <- data.frame(cbind("SITECOORD", "000000", "-21.2500", "165.3000")); #Replace
values with your site's latitude and longitude
colnames(site) <- c('tax', 'ind_id', 'lat', 'lon')
tax <- rbind(site, tax)
###
##Your computer must be able to access the internet in order to perform these analyses.
Specifically it must be able to access www.worldclim.org to download climate data via the
"getData()" function in the dismo packages. If the download does not proceed in R it may
be necessary to download the climate grids manually at
http://biogeo.ucdavis.edu/data/climate/worldclim/1_4/grid/cur/bio_2-5m_bil.zip
####END NOTES
#########################################################
#Everything beyond this point should run without any editing
#########################################################
library(dismo); #Assumes 'dismo' is installed
wd <- getwd();
rawbioclim <- getData('worldclim', var='bio', res=2.5);
#####
print("Welcome to CCC - Community Climate Construction Analysis");
tax$lat <- as.numeric(as.character(tax$lat));
tax$lon <- as.numeric(as.character(tax$lon));
tax <- na.omit(tax);
latmin <- min(tax$lat);
latmax <- max(tax$lat);
```

```
lonmin <- min(tax$lon);
lonmax <- max(tax$lon);
ext = extent((lonmin), (lonmax), (latmin), (latmax));
filter <- c((lonmin), (lonmax), (latmin), (latmax));
geo <- 'Undefined';
tax <- subset(tax, lon > filter[[1]]);
tax <- subset(tax, lon < filter[[2]]);
tax <- subset(tax, lat > filter[[3]]);
tax <- subset(tax, lat < filter[[4]]);
##Some functions to set up first
#extraction gets climate data for every occurrence point
extraction <- function(data, clim, is.bg = F){
	##For now just keep is.bg always FALSE.
	##The procedure below is time consuming and problematic.;
	mat.larr <- data;
	phytoclim <- clim;
	if(is.bg == F){extr.larr <- extract(phytoclim, mat.larr[,4:3], cellnumbers=T);
	extr.larr <- cbind(mat.larr, extr.larr);
	} else {
		extr.larr <- extract(phytoclim, mat.larr[,4:3], cellnumbers=T);
		extr.larr <- cbind(mat.larr, extr.larr);
		obs.bg <- extract(phytoclim, mat.larr[,4:3], buffer = 100000, df = F,
cellnumbers=T);
		obs.bg <- ldply(obs.bg, data.frame);
		m <- data.frame(matrix("bg", ncol = 5, nrow = length(obs.bg[,1])));
		obs.bg <- cbind(m, obs.bg);
		obs.bg <- subset(obs.bg, !obs.bg[,"cell"] %in% extr.larr[,"cells"]);
		obs.bg  <- obs.bg[!duplicated(obs.bg[,"cell"]),];
		obs.bg <- na.omit(obs.bg[,-length(obs.bg[1,])]);
		extr.larr <- obs.bg;
	};
	return(extr.larr);
};
#densform() generates standard PDFs for each taxon/variable
densform <- function(data, clim, bw = "nrd0", bg = "", name = "", boot.n = 1, is.bg = F){
	require(plyr);
	 if(name == ""){
		name = data[2,2];
	 };
	pi = 22/7;
	extr.larr <- data;
	head = 5;
	phytoclim <- clim;
	if(boot.n > 1){		data <- data.frame(extr.larr);
		size <- length(extr.larr[,1]);
		dens.ob <- list();
		for(j in 1:boot.n){				sample.size <- 0.25*size;
			sample <- sample(1:size, sample.size, replace=F);
			data.sam <- data[sample,];
			larr.den <- data.frame();
			larr.den.x <- data.frame();
			larr.den.gauss <- data.frame();
			larr.mean <- data.frame();
			larr.sd <- data.frame();
			eval <- data.frame();
			for(i in 1:length(names(phytoclim))){
				from <- minValue(phytoclim[[i]]);
```

```
                                        to <- maxValue(phytoclim[[i]]);
                                        n = 512;
                                        den <- density(data.sam[,i+head+1], n = n, from = from,
to = to, bw = bw);
                                        mean <- mean(data.sam[,i+head+1]);
                                        sd <- sd(data.sam[,i+head+1]);
                                        if(sd == 0 || is.na(sd) == "TRUE"){
                                                sd = 0.01;
                                        };
                                        for(num in 1:length(den$x)){
                                                eval[num,1] <- ((1/(sqrt(2*pi)*sd)*(2.71828^(-
1*((den$x[num] - mean)^2)/(2*sd^2)))));
                                        };
                                        larr.den[1:n, i] <- den$y;
                                        larr.den.x[1:n, i] <- den$x;
                                        larr.den.gauss[1:n, i] <- eval[,1];
                                        larr.mean[1,i] <- mean;
                                        larr.sd[1,i] <- sd;
                                };
                colnames(larr.den.gauss) <- c(paste(names(phytoclim), "gauss", sep = "."));
                colnames(larr.den) <- c(names(phytoclim));
                colnames(larr.den.x) <- c(paste(names(phytoclim), "x", sep = "."));
                name = data.frame(name);
                larr.mean = data.frame(larr.mean);
                larr.sd = data.frame(larr.sd);
                colnames(name) <- "name";
                fin <- c(larr.den, larr.den.x, larr.den.gauss, larr.mean, larr.sd, name);
                fin <- makeaucone(fin);
                dens.ob[[j]] <- fin;
                };
                return(dens.ob);
        } else {larr.den <- data.frame();
                larr.den.x <- data.frame();
                larr.den.gauss <- data.frame();
                larr.mean <- data.frame();
                larr.sd <- data.frame();
                eval <- data.frame();
                for(i in 1:length(names(phytoclim))){
                        from <- minValue(phytoclim[[i]]);
                        to <- maxValue(phytoclim[[i]]);
                        n = 512;
                        den <- density(extr.larr[,i+head+1], n = n, from = from, to = to,
bw = bw);
                        mean <- mean(extr.larr[,i+head+1]);
                        sd <- sd(extr.larr[,i+head+1]);
                        if(sd == 0 || is.na(sd) == "TRUE"){                             sd = 0.01;
                };
                for(num in 1:length(den$x)){
                        eval[num,1] <- ((1/(sqrt(2*pi)*sd)*(2.71828^(-1*((den$x[num] -
mean)^2)/(2*sd^2)))));
                };
                larr.den[1:n, i] <- den$y;
                larr.den.x[1:n, i] <- den$x;
                larr.den.gauss[1:n, i] <- eval[,1];
                larr.mean[1,i] <- mean;
                larr.sd[1,i] <- sd;
        };
```

```
	colnames(larr.den.gauss) <- c(paste(names(phytoclim), "gauss", sep = "."));
	colnames(larr.mean) <- c(paste(names(phytoclim), "mean", sep = "."));
	colnames(larr.sd) <- c(paste(names(phytoclim), "sd", sep = "."));
	colnames(larr.den) <- c(names(phytoclim));
	colnames(larr.den.x) <- c(paste(names(phytoclim), "x", sep = "."));
	name = data.frame(name);
	larr.mean = data.frame(larr.mean);
	larr.sd = data.frame(larr.sd);
	colnames(name) <- "name";
	fin <- c(larr.den, larr.den.x, larr.den.gauss, larr.mean, larr.sd, name);
	fin <- makeaucone(fin);
	return(fin);
	};
};
#get_optim() takes an object output from the densform function or and_fun or or_fun and
finds optimal values for each PDF
get_optim <- function(dens.ob1, writeOut=F){
	varlist <- names(dens.ob1);
	varlist <- (varlist[1:((length(varlist)-1)/5)]);
	conintkde <- list();
	conintgauss <- list();
	dirconint <- list();
	origk <- list();
	origg <- list();
	means <- list();
	sds <- list();
	for (j in 1:length(varlist)){
		var <- varlist[[j]];
		varx <- paste(var, "x", sep = ".");
		vargauss <- paste(var, "gauss", sep = ".");
		varmean <- paste(var, "mean", sep = ".");
		varsd <- paste(var, "sd", sep = ".");
		cumulkde <- list();
		cumulgauss <- list();
		cikde <- list();
		cigauss <- list();
		runkde <- 0;
		rungauss <- 0;
		to <- max(dens.ob1[[varx]]);
		from <- min(dens.ob1[[varx]]);
		num = length(dens.ob1[[varx]]);
		by = (to - from)/num;
		for (i in 1:length(dens.ob1[[var]])){
			runkde = runkde + (dens.ob1[[var]][i]*by);
			cumulkde[[i]] <- runkde;
		if(i==1){
			if(cumulkde[[i]] >= 0.025){
				cikde[[1]] <- dens.ob1[[varx]][i];
			};
			if(cumulkde[[i]] >= 0.975){
				cikde[[1]] <- dens.ob1[[varx]][i];
				cikde[[2]] <- dens.ob1[[varx]][i];
			};
		} else {
			if(cumulkde[[i-1]] < 0.025 && cumulkde[[i]] >= 0.025){
				cikde[[1]] <- dens.ob1[[varx]][i];
			};
```

```
                        if(cumulkde[[i-1]] < 0.975 && cumulkde[[i]] >= 0.975){
                                cikde[[2]] <- dens.ob1[[varx]][i];
                        };
                };
                rungauss = rungauss + (dens.ob1[[vargauss]][i]*by);
                cumulgauss[[i]] <- rungauss;
                if(i==1){
                        if(cumulgauss[[i]] >= 0.025){
                                cigauss[[1]] <- dens.ob1[[varx]][i];
                        };
                        if(cumulgauss[[i]] >= 0.975){
                                cigauss[[1]] <- dens.ob1[[varx]][i];
                                cigauss[[2]] <- dens.ob1[[varx]][i];
                        };
                } else {
                        if(cumulgauss[[i-1]] < 0.025 && cumulgauss[[i]] >= 0.025){
                                cigauss[[1]] <- dens.ob1[[varx]][i];
                        };
                        if(cumulgauss[[i-1]] < 0.975 && cumulgauss[[i]] >= 0.975){
                                cigauss[[2]] <- dens.ob1[[varx]][i];
                        };
                };
        };
        #maxcukde <- max(as.numeric(cumulkde));
        #cumulkde <- as.numeric(cumulkde)/maxcukde;
        # cikde <- subset(dens.ob1[[varx]], cumulkde >= 0.025 & cumulkde <= 0.975);
        # cikde <- list(c(min(cikde), max(cikde)));
        logkde <- ifelse(dens.ob1[[var]]>0, log(dens.ob1[[var]]*by), -Inf);
        origkde <- subset(dens.ob1[[varx]], logkde >= max(logkde)*1.01);
        origk[[j]] <- c(min(origkde), max(origkde));
         # maxcugauss <- max(as.numeric(cumulgauss));
         # cumulgauss <- as.numeric(cumulgauss)/maxcugauss;
         # cigauss <- subset(dens.ob1[[varx]], cumulgauss >= 0.025 & cumulgauss <=
0.975);
        #cigauss <- list(c(min(cigauss), max(cigauss)));
        loggauss <- ifelse(dens.ob1[[vargauss]]>0, log(dens.ob1[[vargauss]]*by), -Inf);
        origgauss <- subset(dens.ob1[[varx]], loggauss >= max(loggauss)*1.01);
        origg[[j]] <- c(min(origgauss), max(origgauss));
        conintkde[[j]] <- c(cikde[[1]], cikde[[2]]);
        conintgauss[[j]] <- c(cigauss[[1]], cigauss[[2]]);
        dirconint[[j]] <- c((dens.ob1[[varmean]] - 1.96*dens.ob1[[varsd]]),
(dens.ob1[[varmean]]+1.96*dens.ob1[[varsd]]));
        means[[j]] <- dens.ob1[[varmean]];
        sds[[j]] <- dens.ob1[[varsd]];
};
conintkde <- data.frame(conintkde);
conintgauss <- data.frame(conintgauss);
origk <- data.frame(origk);
origg <- data.frame(origg);
dirconint <- data.frame(dirconint);
means <- data.frame(means);
sds <- data.frame(sds);
colnames(conintkde) <- paste(varlist, "cikde", sep = ".");
colnames(conintgauss) <- paste(varlist, "cigauss", sep = ".");
colnames(origk) <- paste(varlist, "origkde", sep = ".");
colnames(origg) <- paste(varlist, "origgauss", sep = ".");
colnames(dirconint) <- paste(varlist, "cidir", sep = ".");
```

```
colnames(means) <- paste(varlist, "mean", sep = ".");
colnames(sds) <- paste(varlist, "sd", sep = ".");
ret <- list(conintkde, conintgauss, origk, origg, dirconint, means, sds);
names(ret) <- c("conintkde", "conintgauss", "origk", "origg", "dirconint", "means", "sds");
return(ret);
};
#makelistarea() corrects area under the curve to be equal to 1 (some lost by pdf
approximation, or for after and_fun or or_fun get new joint pdfs)
makelistarea <- function(list, by){
	do <- sum(list)*by;
	doing <- list/do;
	return(doing);
};
makeaucone <- function(dens.ob1, var){
	var <- names(dens.ob1);
	var <- (var[1:((length(var)-1)/5)]);
	for(i in 1:length(var)){
		varnow <- var[[i]];
		varx <- paste(var[[i]], "x", sep = ".");
		gauss <- paste(var[[i]], "gauss", sep = ".");
		to <- max(dens.ob1[[varx]]);
		from <- min(dens.ob1[[varx]]);
		num <- length(dens.ob1[[varx]]);
		by = (to - from)/num;
		do <- sum(dens.ob1[[varnow]])*by;
		do.gauss <- sum(dens.ob1[[gauss]])*by;
		dens.ob1[[varnow]] <- dens.ob1[[varnow]]/do;
		dens.ob1[[gauss]] <- dens.ob1[[gauss]]/do.gauss;
	};
	return(dens.ob1);
};
#or_fun P(A OR B) = P(A) + P(B)
or_fun <- function(dens.oblist){
	varlist <- names(dens.oblist[[1]]);
	varlist <- (varlist[1:((length(varlist)-1)/5)]);
	field <- list();
	gfield <- list();
	xfield <- list();
	meanadjust <- list();
	variances <- list();
	name = "ADDITION";
	for (n in 1:length(varlist)){
		var = varlist[n];
		varx <- paste(var, "x", sep = ".");
		vargauss <- paste(var, "gauss", sep = ".");
		varmean <- paste(var, "mean", sep = ".");
		varsd <- paste(var, "sd", sep = ".");
		meanlist <- list();
		sdlist <- list();
		dens.obcurr <- dens.oblist[[1]];
		to <- max(dens.obcurr[[varx]]);
		from <- min(dens.obcurr[[varx]]);
		num = length(dens.obcurr[[varx]]);
		by = (to - from)/num;
		meanlist[[1]] <- as.numeric(dens.obcurr[[varmean]]);
		sdlist[[1]] <- as.numeric(dens.obcurr[[varsd]])^2;
```

```
                prod <- as.numeric(dens.obcurr[[var]]);
                prod.gauss <- as.numeric(dens.obcurr[[vargauss]]);
                for(i in 2:length(dens.oblist)){dens.obnow <- dens.oblist[[i]];
                        prod <- prod + (as.numeric(dens.obnow[[var]]));
                        prod.gauss <- prod.gauss + (as.numeric(dens.obnow[[vargauss]]));
                        prod.area <- sum(prod)*by;
                        prod <- prod / prod.area;
                        prod.gauss.area <- sum(prod.gauss)*by;
                        prod.gauss <- prod.gauss / prod.gauss.area;
                        meanlist[[i]] <- as.numeric(dens.obnow[[varmean]]);
                        sdlist[[i]] <- as.numeric(dens.obnow[[varsd]])^2;
                };
                prod.area <- sum(prod)*by;
                prod <- prod / prod.area;
                prod.gauss.area <- sum(prod.gauss)*by;
                prod.gauss <- prod.gauss / prod.gauss.area;
                field[[n]] <- prod;
                gfield[[n]] <- prod.gauss;
                xfield[[n]] <- dens.obcurr[[varx]];
                meanadjust[[n]] <- as.numeric(meanlist)/as.numeric(sdlist);
                variances[[n]] <- 1/as.numeric(sdlist);
        };
        meansum <- lapply(meanadjust, sum);
        varisum <- lapply(variances, sum);
        wmeans <- mapply("/", meansum, varisum);
        wsd <- mapply("/", 1, varisum);
        wsd <- lapply(wsd, sqrt);
        field <- data.frame(field);
        gfield <- data.frame(gfield);
        xfield <- data.frame(xfield);
        colnames(field) <- (varlist);
        colnames(gfield) <- paste(varlist, "gauss", sep = ".");
        names(wmeans) <- paste(varlist, "mean", sep = ".");
        names(wsd) <- paste(varlist, "sd", sep = ".");
        colnames(xfield) <- (paste(varlist, "x", sep = "."));
        name = data.frame(name);


        colnames(name) <- "name";
        fin <- c(field, xfield, gfield, wmeans, wsd, name);
        fin <- makeaucone(fin);
        return(fin);
};
#and_fun() P(A and B) = P(A) * P(B) or log(P(A)) + log(P(B))
and_fun <- function(dens.oblist){
        dens.oblist <- scramble(dens.oblist);
        varlist <- names(dens.oblist[[1]]);
        varlist <- (varlist[1:((length(varlist)-1)/5)]) ;
        field <- list();
        gfield <- list();
        xfield <- list();
        meanadjust <- list();
        variances <- list();
        name = "PRODUCT";
        for (n in 1:length(varlist)){var = varlist[n];
                        varx <- paste(var, "x", sep = ".");
                        vargauss <- paste(var, "gauss", sep = ".");
                        varmean <- paste(var, "mean", sep = ".");
```

```
                        varsd <- paste(var, "sd", sep = ".");
                        meanlist <- list();
                        sdlist <- list();
                        dens.obcurr <- dens.oblist[[1]];
                        to <- max(dens.obcurr[[varx]]);
                        from <- min(dens.obcurr[[varx]]);
                        num = length(dens.obcurr[[varx]]);
                        by = (to - from)/num;
                        meanlist[[1]] <- as.numeric(dens.obcurr[[varmean]]);
                        sdlist[[1]] <- as.numeric(dens.obcurr[[varsd]])^2;
                        prod <- as.numeric(dens.obcurr[[var]]);
                        prod <- prod*by;
                        prod.gauss <- as.numeric(dens.obcurr[[vargauss]])*by;
                        for(i in 2:length(dens.oblist)){dens.obnow <- dens.oblist[[i]];
                                prod <- prod * (as.numeric(dens.obnow[[var]])*by);
                                prod.area <- sum(prod)*by;
                                 prod <- prod / prod.area;
                                prod.gauss <- prod.gauss *
(as.numeric(dens.obnow[[vargauss]])*by);
                                 prod.gauss.area <- sum(prod.gauss)*by;
                                 prod.gauss <- prod.gauss / prod.gauss.area;
                                meanlist[[i]] <- as.numeric(dens.obnow[[varmean]]);
                                sdlist[[i]] <- as.numeric(dens.obnow[[varsd]])^2;
                        };
                        prod.area <- sum(prod)*by;
                        prod <- prod / prod.area;
                        prod.gauss.area <- sum(prod.gauss)*by;
                        prod.gauss <- prod.gauss / prod.gauss.area;
                        field[[n]] <- prod;
                        gfield[[n]] <- prod.gauss;
                        xfield[[n]] <- dens.obcurr[[varx]];
                        meanadjust[[n]] <- as.numeric(meanlist)/as.numeric(sdlist);
                        variances[[n]] <- 1/as.numeric(sdlist);
                };
        meansum <- lapply(meanadjust, sum);
        varisum <- lapply(variances, sum);
        wmeans <- mapply("/", meansum, varisum);
        wsd <- mapply("/", 1, varisum);
        wsd <- lapply(wsd, sqrt);
        field <- data.frame(field);
        gfield <- data.frame(gfield);
        xfield <- data.frame(xfield);
        colnames(field) <- (varlist);
        colnames(gfield) <- paste(varlist, "gauss", sep = ".");
        names(wmeans) <- paste(varlist, "mean", sep = ".");
        names(wsd) <- paste(varlist, "sd", sep = ".");
        colnames(xfield) <- (paste(varlist, "x", sep = "."));
        name = data.frame(name);
        colnames(name) <- "name";
        fin <- c(field, xfield, gfield, wmeans, wsd, name);
        fin <- makeaucone(fin);
        return(fin);
};
#scramble() reorders pdfs. No real reason to do this as order does not matter for the
operations being done here.
scramble <- function(x, k=3) {
        x.s <- seq_along(x);
```

```
	y.s <- sample(x.s);
	idx <- unlist(split(y.s, (match(y.s, x.s)-1) %/% k), use.names = FALSE);
	x[idx];
};
#densplot() Plot PDF
densplot <- function(dens.ob, var, col = sample(colours()), type = "") {
	varx <- paste(var, "x", sep = ".");
	par(mar= c(5,4,4,4) + 0.3);
	tempvarlist <- c("bio1", "bio2", "bio3", "bio4", "bio5", "bio6", "bio7", "bio8", "bio9",
"bio10", "bio11", "MAT", "MaximumT", "MinimumT");
	if(var %in% tempvarlist){by = 10}else{by = 1};
	var <- paste(var, type, sep = "");
	plot(dens.ob[[varx]]/by, dens.ob[[var]], xlab = "", ylab = "", ylim = c(0,
4*max(dens.ob[[var]])), type = "l", lwd = 3, col = col, frame.plot=F, axes = F);
	axis(side = 2, at = pretty(c(0, 4*max(dens.ob[[var]]))));
	axis(side = 1, at = pretty(range(dens.ob[[varx]]/by)));
	mtext(var, side = 1, line =3);
	mtext("Kernel Density Estimation", side = 2, line = 3);
};
#addplot() adds PDF plot to already open plot
addplot <- function(dens.ob, var, col = sample(colours()), type = "") {
	varx <- paste(var, "x", sep = ".");
	tempvarlist <- c("bio1", "bio2", "bio3", "bio4", "bio5", "bio6", "bio7", "bio8", "bio9",
"bio10", "bio11", "MAT", "MaximumT", "MinimumT");
	if(var %in% tempvarlist){by = 10}else{by = 1};
	var <- paste(var, type, sep = "");
	points(dens.ob[[varx]]/by, dens.ob[[var]], type = "l", lwd = 3, col = col);
};
#multiplot() a wrapper for densplot and addplot together given a list of PDFs
multiplot <- function(arr.dens.ob, var, col = colours(length(arr.dens.ob)), type = ""){
	varx <- paste(var, "x", sep = ".");
	current <- arr.dens.ob[[1]];
	densplot(current, var, col[1], type = type);
	max.x.hold = list(max(current[[varx]]));
	max.y.hold = list(max(current[[var]]));
	names.hold = as.character(current[["name"]]);
	for(i in 2:length(arr.dens.ob)){current <- arr.dens.ob[[i]];
		addplot(current, var, col[i], type = type);
		max.x.hold = c(max.x.hold, max(current[[varx]]));
		max.y.hold = c(max.y.hold, max(current[[var]]));
		names.hold = c(names.hold, as.character(current[["name"]]));
	};
	max.x <- mean(as.numeric(as.character(max.x.hold)));
	max.y <- mean(as.numeric(as.character(max.y.hold)));
	legend("topleft", legend = as.character(names.hold), lty=1, lwd=2, cex=0.8, col =
col, box.col=NA);
};
##BODY OF CODE
tax <- data.frame(tax);
dens.list <- list();
ex <- extraction(tax, rawbioclim);
site.ex <- "NOSITE";
site.coord = 0;
if(tax[1,2] == "SITECOORD"){ site.coord <- tax[1,];
	tax <- subset(tax, tax != "SITECOORD");
	site.ex <- ex[1,];
 print(site.ex);
```

```
 };
tax.list <- unique(tax$tax);
tax.list <- na.omit(tax.list);
for(i in 1:length(tax.list)){         print("i is:");
          print(i);
          s.ex <- subset(ex, ex$tax == tax.list[[i]]);
                    s.ex <- s.ex[!duplicated(s.ex[,"cells"]),];
          s.ex <- na.omit(s.ex);
                    dens.list[[i]] <- densform(s.ex, rawbioclim, name = tax.list[[i]], boot.n=1);
                    len <- length(dens.list[[i]]);
          if(len <= 1) {               print("OTHER");
                    dens.list[[i]] <- NULL;
                              };
};
anding <- and_fun(dens.list);
optima <- get_optim(anding);
print(optima);
MLinfer.comp = data.frame(names(rawbioclim));
for(i in 1:length(names(rawbioclim))){  MLinfer.comp[i,2] = site.ex[i+6];
          MLinfer.comp[i,3] = optima[["origk"]][[i]][1];
          MLinfer.comp[i,4] = optima[["origk"]][[i]][2];
          MLinfer.comp[i,5] = as.numeric(MLinfer.comp[i,3])-as.numeric(MLinfer.comp[i,2]);
          MLinfer.comp[i,6] = as.numeric(MLinfer.comp[i,4])-as.numeric(MLinfer.comp[i,2]);
};
colnames(MLinfer.comp) = c("climate_variable", "site_value", "MLinfer_min",
"MLinfer_max", "min_resid", "max_resid");
bincountML.comp = data.frame(names(rawbioclim));
for(i in 1:length(names(rawbioclim))){  bincountML.comp[i,2] = site.ex[i+6];
          bincountML.comp[i,3] = optima[["origg"]][[i]][1];
          bincountML.comp[i,4] = optima[["origg"]][[i]][2];
          bincountML.comp[i,5] = as.numeric(bincountML.comp[i,3])-
as.numeric(bincountML.comp[i,2]);
          bincountML.comp[i,6] = as.numeric(bincountML.comp[i,4])-
as.numeric(bincountML.comp[i,2]);
};
colnames(bincountML.comp) = c("climate_variable", "site_value", "bincountML_min",
"bincountML_max", "min_resid", "max_resid");
conintgauss.comp = data.frame(names(rawbioclim));
for(i in 1:length(names(rawbioclim))){  conintgauss.comp[i,2] = site.ex[i+6];
          conintgauss.comp[i,3] = optima[["conintgauss"]][[i]][1];
          conintgauss.comp[i,4] = optima[["conintgauss"]][[i]][2];
          conintgauss.comp[i,5] = as.numeric(conintgauss.comp[i,3])-
as.numeric(conintgauss.comp[i,2]);
          conintgauss.comp[i,6] = as.numeric(conintgauss.comp[i,4])-
as.numeric(conintgauss.comp[i,2]);
};
colnames(conintgauss.comp) = c("climate_variable", "site_value", "conintgauss_min",
"conintgauss_max", "min_resid", "max_resid");
conintkde.comp = data.frame(names(rawbioclim));
for(i in 1:length(names(rawbioclim))){  conintkde.comp[i,2] = site.ex[i+6];
          conintkde.comp[i,3] = optima[["conintkde"]][[i]][1];
          conintkde.comp[i,4] = optima[["conintkde"]][[i]][2];
          conintkde.comp[i,5] = as.numeric(conintkde.comp[i,3])-
as.numeric(conintkde.comp[i,2]);
          conintkde.comp[i,6] = as.numeric(conintkde.comp[i,4])-
as.numeric(conintkde.comp[i,2]);
};
```

```
colnames(conintkde.comp) = c("climate_variable", "site_value", "conintkde_min",
"conintkde_max", "min_resid", "max_resid");
dirconint.comp = data.frame(names(rawbioclim));
for(i in 1:length(names(rawbioclim))){  dirconint.comp[i,2] = site.ex[i+6];
        dirconint.comp[i,3] = optima[["dirconint"]][[i]][1];
        dirconint.comp[i,4] = optima[["dirconint"]][[i]][2];
        dirconint.comp[i,5] = as.numeric(dirconint.comp[i,3])-
as.numeric(dirconint.comp[i,2]);
        dirconint.comp[i,6] = as.numeric(dirconint.comp[i,4])-
as.numeric(dirconint.comp[i,2]);
};
colnames(dirconint.comp) = c("climate_variable", "site_value", "dirconint_min",
"dirconint_max", "min_resid", "max_resid");
means.comp = data.frame(names(rawbioclim));
for(i in 1:length(names(rawbioclim))){  means.comp[i,2] = site.ex[i+6];
        means.comp[i,3] = optima[["means"]][[i]][1];
        means.comp[i,4] = optima[["means"]][[i]][1];
        means.comp[i,5] = as.numeric(means.comp[i,3])-as.numeric(means.comp[i,2]);
        means.comp[i,6] = as.numeric(means.comp[i,4])-as.numeric(means.comp[i,2]);
};
colnames(means.comp) = c("climate_variable", "site_value", "means_min", "means_max",
"min_resid", "max_resid");
tax_used <- ex[,1:4];
#Write results and metadata to the "est.tab" file in the local directory
write.table("sitecoord", file="est.tab", append=T, sep="          ", quote=F, row.names=F);
write.table(site.coord, file="est.tab", append=T, sep=" ", quote=F, row.names=F);
write.table("TaxNum", file="est.tab", append =T, sep = "          ", quote=F, row.names=F);
write.table(length(tax.list), file="est.tab", append =T, sep = "   ", quote=F, row.names=F);
write.table("TaxList", file="est.tab", append =T, sep = "          ", quote=F, row.names=F);
write.table(tax.list, file="est.tab", append =T, sep = "   ", quote=F, row.names=F);
write.table(">>>>MLinfer", file="est.tab", append=T, sep="   ", quote=F, row.names=F,
col.names=F);
write.table(MLinfer.comp, file="est.tab", append=T, sep="       ", quote=F, row.names=F,
col.names=T);
write.table(">>>>bincountML", file="est.tab", append=T, sep="         ", quote=F,
row.names=F, col.names=F);
write.table(bincountML.comp, file="est.tab", append=T, sep="", quote=F, row.names=F,
col.names=T);
write.table(">>>>conintgauss", file="est.tab", append=T, sep="         ", quote=F,
row.names=F, col.names=F);
write.table(conintgauss.comp, file="est.tab", append=T, sep="", quote=F, row.names=F,
col.names=T);
write.table(">>>>conintkde", file="est.tab", append=T, sep="          ", quote=F,
row.names=F, col.names=F);
write.table(conintkde.comp, file="est.tab", append=T, sep="    ", quote=F, row.names=F,
col.names=T);
write.table(">>>>dirconint", file="est.tab", append=T, sep="  ", quote=F, row.names=F,
col.names=F);
write.table(dirconint.comp, file="est.tab", append=T, sep="     ", quote=F, row.names=F,
col.names=T);
write.table(">>>>means", file="est.tab", append=T, sep="     ", quote=F, row.names=F,
col.names=F);
write.table(means.comp, file="est.tab", append=T, sep="         ", quote=F, row.names=F,
col.names=T);
removeTmpFiles();
 ##Removes temporary raster files
q("no");
```

##R WILL EXIT HERE. Your results will be in the “est.tab” file in the local directory

Table S2.2: CRACLE KDE Mean Anomalies

	KDE 0	KDE 20	KDE 35	KDE 50	KDE 65	KDE 90	KDE 100
MAT	1.49	1.66	1.83	1.94	2.11	2.43	2.65
MaximumT	1.32	1.50	1.70	1.87	2.07	2.62	3.09
MinimumT	2.08	2.28	2.40	2.65	2.90	3.67	4.14
tempbalance	1.83	2.02	2.23	2.40	2.61	2.98	3.27
diurnal	0.96	1.01	1.08	1.10	1.18	1.46	1.69
GSL	0.86	0.98	1.04	1.14	1.23	1.43	1.58
MAP	267.75	282.81	295.71	318.95	344.24	402.31	475.31
GSPREC	275.79	287.45	299.79	321.75	348.67	396.19	468.33
wbalann	289.85	341.44	366.98	400.88	439.74	515.83	565.01
maxwbal	4.15	4.55	4.82	5.18	5.54	6.22	6.87
minwbal	2.22	2.57	2.81	2.90	3.11	3.41	3.64
wet_mo	1.06	1.11	1.17	1.20	1.31	1.42	1.57
wet_mo_mean	21.12	24.65	26.51	28.71	31.06	34.56	37.77
wet_sum	23.68	28.02	30.12	32.71	36.13	42.10	45.77
dry_mo	1.06	1.11	1.17	1.20	1.31	1.41	1.57
dry_mo_mean	9.33	10.01	10.43	11.38	12.66	14.65	16.29
dry_sum	84.57	92.91	100.32	108.95	119.82	135.10	146.88
X3DryP	55.82	65.44	72.37	78.41	82.79	91.25	96.98
X3WetP	119.43	130.95	136.26	144.41	151.54	165.99	181.97
DRLEN	0.82	0.88	0.94	1.04	1.14	1.27	1.38
DRSEV	12.66	13.86	14.76	15.93	16.78	20.51	23.35
WINTERLEN	0.32	0.32	0.34	0.34	0.36	0.47	0.55

Table S2.3: CRACLE Gaussian Mean Anomalies

	Gaussian 0	Gaussian 20	Gaussian 35	Gaussian 50	Gaussian 65	Gaussian 90	Gaussian 100
MAT	1.44	1.53	1.59	1.66	1.72	1.90	2.03
MaximumT	1.38	1.53	1.66	1.81	1.95	2.34	2.67
MinimumT	2.06	2.18	2.26	2.37	2.49	2.82	3.07
tempbalance	1.76	1.87	1.94	2.01	2.08	2.32	2.46
diurnal	0.89	0.94	0.99	1.07	1.14	1.35	1.55
GSL	0.86	0.92	0.98	1.03	1.12	1.31	1.51
MAP	295.16	327.76	342.38	369.18	397.60	441.70	481.63
GSPREC	304.84	339.56	355.07	381.67	412.11	459.21	500.39
wbalann	317.54	347.62	365.74	390.61	419.33	463.21	498.42
maxwbal	4.32	4.67	4.89	5.23	5.50	6.04	6.48
minwbal	2.26	2.43	2.56	2.69	2.85	3.16	3.36
wet_mo	1.16	1.24	1.32	1.36	1.46	1.62	1.75
wet_mo_mean	23.01	25.13	26.48	28.71	30.37	33.85	36.19
wet_sum	25.63	28.11	29.71	31.78	34.04	37.65	40.07
dry_mo	1.16	1.24	1.32	1.36	1.46	1.62	1.75
dry_mo_mean	10.23	10.94	11.45	11.88	13.00	14.73	16.09

dry_sum	74.94	80.00	85.45	90.31	97.90	112.10	125.33
X3DryP	58.29	64.60	67.50	71.37	76.81	83.51	89.71
X3WetP	119.35	127.69	133.42	143.32	150.59	163.91	176.19
DRLEN	0.73	0.77	0.85	0.91	0.99	1.16	1.35
DRSEV	10.33	10.81	11.63	12.47	13.34	15.33	17.57
WINTERLEN	0.24	0.25	0.26	0.26	0.27	0.32	0.36

Table S2.4: CRACLE KDE median anomaly

	KDE 0	KDE 20	KDE 35	KDE 50	KDE 65	KDE 90	KDE 100
MAT	1.05	1.22	1.41	1.42	1.52	1.64	1.63
MaximumT	0.84	0.96	1.23	1.31	1.42	1.80	2.13
MinimumT	1.32	1.65	1.70	1.94	2.16	2.92	3.40
tempbalance	12.98	14.72	16.82	17.38	19.10	20.11	20.72
diurnal	0.56	0.58	0.65	0.66	0.74	1.00	1.08
GSL	0.15	0.13	0.31	0.54	0.62	0.96	1.01
MAP	83.62	94.53	107.82	121.32	152.92	207.54	264.60
GSPREC	83.94	92.54	102.05	126.44	154.94	190.34	237.22
wbalann	140.94	159.86	177.42	199.47	239.18	300.37	326.64
maxwbal	1.62	1.89	2.06	2.16	2.41	2.93	3.34
minwbal	1.19	1.30	1.49	1.72	1.92	2.22	2.64
wet_mo	0.88	0.93	0.95	0.95	1.00	1.01	1.17
wet_mo_mean	9.24	12.06	12.03	13.44	15.65	18.33	18.99
wet_sum	8.35	9.85	11.13	12.68	14.49	17.46	18.47
dry_mo	0.88	0.93	0.95	0.95	1.00	1.01	1.17
dry_mo_mean	5.06	5.28	6.27	7.05	7.86	9.86	11.52
dry_sum	45.15	51.26	57.36	65.19	71.17	82.62	93.72
X3DryP	15.69	18.96	23.14	25.84	29.16	37.82	40.51
X3WetP	50.98	56.45	65.35	73.99	83.53	95.59	102.90
DRLEN	0.07	0.06	0.06	0.06	0.07	0.07	0.07
DRSEV	0.35	0.35	0.35	0.35	0.35	0.60	0.70
WINTERLEN	0.02	0.01	0.01	0.01	0.00	0.00	0.00

Table S2.5: CRACLE Gaussian median anomaly

	Gaussian 0	Gaussian 20	Gaussian 35	Gaussian 50	Gaussian 65	Gaussian 90	Gaussian 100
MAT	1.10	1.11	1.12	1.13	1.22	1.33	1.46
MaximumT	0.94	1.03	1.13	1.19	1.39	1.73	2.07
MinimumT	1.54	1.65	1.72	1.77	1.97	2.20	2.31
tempbalance	13.59	13.44	13.98	14.15	14.93	16.06	17.01
diurnal	0.61	0.58	0.64	0.77	0.81	0.95	1.18
GSL	0.59	0.66	0.75	0.81	0.91	1.05	1.24
MAP	146.84	180.38	179.88	192.15	224.99	246.40	287.80

GSPREC	156.18	192.89	192.66	212.15	242.09	290.35	315.32
wbalann	169.32	185.75	203.52	226.06	240.70	290.17	314.83
maxwbal	2.11	2.31	2.23	2.66	2.96	3.35	3.89
minwbal	1.35	1.43	1.67	1.72	1.77	2.11	2.39
wet_mo	0.92	1.00	1.04	1.11	1.20	1.40	1.50
wet_mo_mean	12.65	14.25	15.58	17.03	16.69	18.97	21.36
wet_sum	11.78	14.24	14.80	14.32	17.59	19.49	20.05
dry_mo	0.92	1.00	1.04	1.11	1.20	1.40	1.50
dry_mo_mean	7.49	8.71	8.63	9.17	9.91	11.80	13.33
dry_sum	50.44	55.25	59.51	64.76	74.21	88.47	105.40
X3DryP	31.01	33.88	33.80	39.03	42.06	46.76	58.82
X3WetP	61.18	68.99	68.58	79.36	88.09	92.77	102.69
DRLEN	0.39	0.41	0.53	0.60	0.74	0.90	1.02
DRSEV	2.61	2.79	3.17	3.86	5.30	7.34	8.08
WINTERLEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table S2.6: CRACLE KDE r2

	KDE 0	KDE 20	KDE 35	KDE 50	KDE 65	KDE 90	KDE 100
MAT	0.90	0.87	0.85	0.83	0.80	0.73	0.67
MaximumT	0.82	0.77	0.72	0.67	0.60	0.40	0.21
MinimumT	0.93	0.91	0.90	0.89	0.87	0.80	0.75
tempbalance	0.90	0.88	0.86	0.83	0.80	0.74	0.68
diurnal	0.70	0.68	0.64	0.63	0.59	0.40	0.18
GSL	0.85	0.81	0.79	0.76	0.74	0.67	0.62
MAP	0.75	0.74	0.73	0.71	0.70	0.64	0.55
GSPREC	0.77	0.76	0.75	0.73	0.72	0.67	0.58
wbalann	0.72	0.63	0.59	0.55	0.50	0.37	0.29
maxwbal	0.64	0.58	0.55	0.50	0.46	0.39	0.30
minwbal	0.66	0.55	0.51	0.50	0.46	0.41	0.37
wet_mo	0.76	0.74	0.71	0.70	0.66	0.63	0.57
wet_mo_mean	0.73	0.62	0.59	0.52	0.45	0.40	0.30
wet_sum	0.62	0.53	0.49	0.43	0.35	0.19	0.09
dry_mo	0.76	0.74	0.71	0.70	0.66	0.63	0.57
dry_mo_mean	0.73	0.70	0.68	0.63	0.56	0.44	0.36
dry_sum	0.74	0.70	0.66	0.61	0.53	0.42	0.32
X3DryP	0.47	0.36	0.29	0.23	0.19	0.09	0.05
X3WetP	0.74	0.71	0.70	0.68	0.66	0.64	0.59
DRLEN	0.78	0.74	0.71	0.65	0.60	0.52	0.45
DRSEV	0.66	0.60	0.57	0.53	0.49	0.26	0.07
WINTERLEN	0.85	0.84	0.83	0.83	0.81	0.70	0.60

Table S2.7: CRACLE Gaussian r2

	Gaussian 0	Gaussian 20	Gaussian 35	Gaussian 50	Gaussian 65	Gaussian 90	Gaussian 100
MAT	0.92	0.90	0.89	0.88	0.87	0.84	0.82
MaximumT	0.84	0.79	0.76	0.72	0.69	0.58	0.47
MinimumT	0.93	0.93	0.92	0.91	0.90	0.88	0.86
tempbalance	0.92	0.90	0.90	0.89	0.88	0.85	0.83
diurnal	0.77	0.74	0.72	0.68	0.64	0.53	0.40
GSL	0.89	0.88	0.86	0.85	0.83	0.80	0.75
MAP	0.82	0.78	0.76	0.70	0.68	0.62	0.57
GSPREC	0.83	0.79	0.77	0.72	0.69	0.63	0.59
wbalann	0.77	0.72	0.70	0.63	0.60	0.52	0.47
maxwbal	0.71	0.67	0.64	0.59	0.56	0.50	0.46
minwbal	0.72	0.67	0.64	0.61	0.58	0.51	0.47
wet_mo	0.78	0.76	0.74	0.73	0.69	0.63	0.59
wet_mo_mean	0.76	0.72	0.69	0.62	0.59	0.51	0.46
wet_sum	0.74	0.67	0.64	0.56	0.52	0.43	0.37
dry_mo	0.78	0.76	0.74	0.73	0.69	0.63	0.59
dry_mo_mean	0.77	0.74	0.73	0.71	0.66	0.58	0.52
dry_sum	0.82	0.80	0.78	0.76	0.73	0.68	0.62
X3DryP	0.66	0.57	0.53	0.49	0.43	0.35	0.31
X3WetP	0.81	0.79	0.77	0.72	0.70	0.66	0.63
DRLEN	0.85	0.84	0.82	0.80	0.77	0.72	0.65
DRSEV	0.81	0.79	0.76	0.74	0.72	0.65	0.58
WINTERLEN	0.93	0.92	0.92	0.91	0.90	0.87	0.84

APPENDIX 3:

SUPPLEMENTAL METHODS AND MATERIALS:

CRACLE –

Estimation of climate based on species coexistence and modern species distributions via the Climate Reconstruction Analysis Using Coexistence Likelihood Estimation (CRACLE) protocol (Harbert and Nixon, 2015). This method generates parametric (normal Gaussian) and non-parametric (Gaussian Kernel Density Estimation) probability functions for the occurrence of a species along a dimension of climate (e.g., average annual temperature). The joint likelihood function for all co-occurring species is then calculated as the product (or sum-log-likelihood) of these species functions. The maximum of the joint likelihood curve is taken to be the most probable climate value given the association of species and their individual association with climate.

For this study, CRACLE was implemented in the previously described (Harbert and Nixon, 2015) manner with only two changes (See R script in Appendix S2.1). First, to optimize the Kernel Density Estimation procedure Silverman's Rule was applied to select the near optimal bandwidth (Silverman, 1986) rather than using pre-defined bandwidth for each variable (Harbert and Nixon, 2015). Second, all taxa identified to at least genus were included for this study, whereas in the original implementation only species identifications were included. This was done to expand the sample of middens that could be used to cover a broader time period and with higher spatial resolution.

USGS-NOAA Packrat Midden Database –

The work of countless researchers (notably the volume: Betancourt et al., 1990) in recent decades to locate, catalog, and identify fossil plant remains in the middens of *Neotoma* spp. rats is now largely curated by the United States Geological Survey in the form of the USGS-NOAA Packrat Midden Database (available online: http://geochange.er.usgs.gov/midden/search.html). This database provides access to taxonomic identification, locality information and georeferencing, and carbon dating data. Fifty additional sites were gleaned from the work of Holmgren et al. (2014) at Guadalupe Canyon in the Lower Colorado River Basin in Northern Mexico.

For this study, these data were filtered to identify dated and georeferenced middens with fossils identified to 10 or more modern plant taxa (genus or species). Georeferencing precision was filtered so that only sites with precision of less than one 0.5 degrees in both latitude and longitude, suggesting that these sites were more carefully georeferenced (2.5 arcmin WorldClim grid). Dated layers from the same locality were treated as separate units to allow estimates for middens to change through time as their record of vegetation changes.

Midden Age Chronology and Paleoclimate Timeline --

The atmospheric 14C/12C ratio is not stable through time, therefore radiocarbon derived dates must be calibrated to calendar years in order to correctly align these dates with events or data given in calendar years. Radiocarbon dates for the packrat middens were calibrated using the Northern Hemisphere IntCal13 calibration curve (Reimer, 2013), and the calibration tools available in the “Bchron” software library (Parnell, 2015) in R (R Core Team, 2016).

Using the calibrated midden age confidence intervals, time bins were established at intervals of 200 years from -50,000 to the present the mean climate values were calculate including all middens with a confidence interval overlapping the 200 year bin.

Modern climate model –

Climate data are taken from the downscaled 2.5-arcminute resolution (~0.041667 degrees) WorldClim model grid (Hijmans et al., 2005).WorldClim is a high-resolution continuous grid of interpolated climate data for the world’s land areas derived from >40 000 weather stations around the world.

Climate Variables –

One of the major advantages of the CRACLE method for inferring paleoclimate is that the general protocol is flexible and can be applied to a wide variety of climatic parameters. For this study quantitative reconstructions of the variables mean annual temperature (MAT), maximum temperature, minimum temperature, mean precipitation, mean water balance (potential evapotranspiration + mean precipitation), and winter length (months with mean temperature < 5°C) have been focused on.

To better capture nuances of climate relating to drought the potential evapotranspiration (PET) and water balance (precipitation – PET) was calculated using monthly values for temperature and precipitation from the WorldClim model (Hijmans et al., 2005). The Thornwaite Equation (Thornwaite, 1948) was chosen as a suitable model of potential evapotranspiration that relied on data available at the scale of WorldClim including: 1) monthly average temperatures in degrees Celcius, 2) daylength in hours (calculated from the latitude and month), 3) the number of days in each month.

Modern test sites – Validation and alignment procedure

To gauge the relative difference of CRACLE from the WorldClim 93 modern data vegetation surveys were collected from the study area (125W, 100W, 23N, 49N). CRACLE was performed on these sites in the same way as for the midden samples. Relative errors for each variable were calculated. To account for prediction biases in the CRACLE method the direction and magnitude of relative error were used as empirically derived correction factors to adjust the paleoclimate reconstructions appropriately (Table S3.1). An example of how this works is shown in Fig.

S3.1.

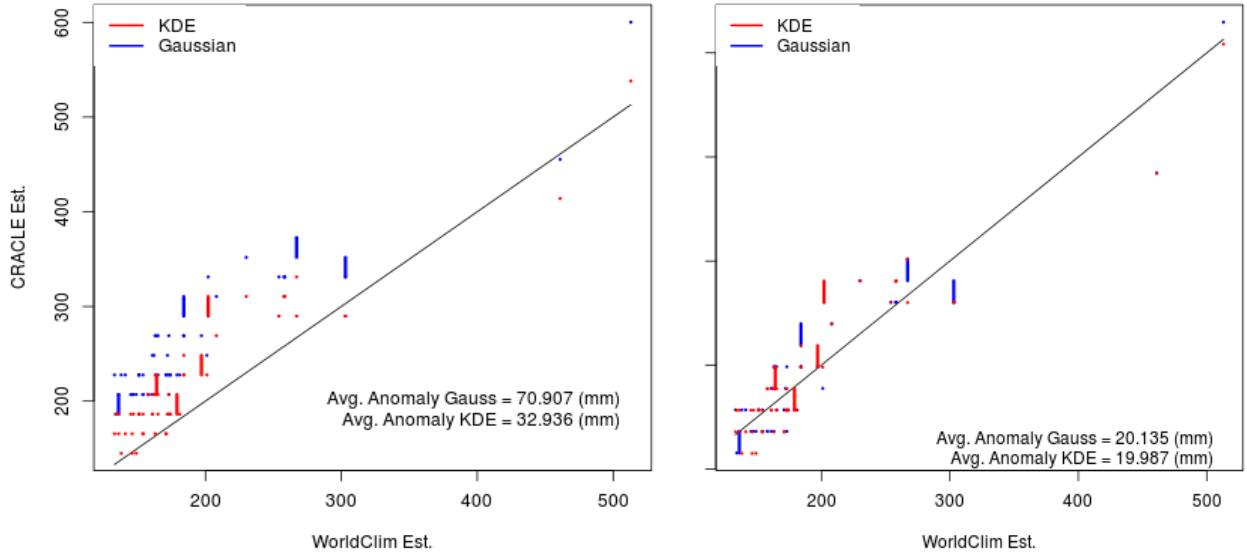

Figure S3.1. Example Modern calibration Mean Annual Precipitation example. Modern vegetation surveys from 55 localities across the study area were analyzed using CRACLE to evaluate relative error made by the model in this region (A). To correct for biased results (e.g., the consistent overestimation of MAP in this region by CRACLE) the average error was calculated for the CRACLE results, then the results were uniformly adjusted up or down by that amount and the error was recalculated (B). Average anomaly values reported are equivalent to the average absolute value of the difference between the CRACLE estimate and the WorldClim estimate for each site.

Appendix Table S3.1: Western North American Correction Factors

var	kdeerr	gausserr
bio10_est	-0.524	-0.614
bio11_est	1.284	0.852
bio12_est	29.682	70.703
bio13_est	5.016	12.69
bio14_est	-2.389	-0.641
bio15_est	8.246	12.729
bio16_est	11.11	34.544
bio17_est	-5.771	-3.007
bio18_est	-5.264	1.461
bio19_est	-1.635	19.876
bio1_est	0.752	0.245
bio2_est	0.447	-0.037
bio3_est	0.085	0.127
bio4_est	-25.703	-38.028

bio5_est	-0.461	-0.583
bio6_est	1.862	1.156
bio7_est	-1.176	-1.166
bio8_est	1.793	2.922
bio9_est	2.726	1.698
diurnal_est	0.044	-0.004
DRLEN_est	-1.251	-1.118
DRSEV_est	-24.672	-18.588
dry_mo_est	-1.068	-0.703
dry_mo_mean_est	-0.997	-0.244
dry_sum_est	95.933	68.87
GSL_est	0.999	1.171
GSPREC_est	3.635	51.242
MAP_est	29.682	70.703
MAT_est	0.752	0.245
MaximumT_est	-0.461	-0.583
maxtemp_est	-0.461	-0.583
maxwbal_est	1.015	6.548
MinimumT_est	1.862	1.156
mintemp_est	1.862	1.156
minwbal_est	12.825	6.877
tempbalance_est	9.887	5.395
wbalann_est	128.514	93.628
wet_mo_est	1.068	0.702
wet_mo_mean_est	0.337	4.414
wet_sum_est	-5.14	28.295
WINTERLEN_est	-1.201	-0.522
X3DryP_est	-5.771	-3.007
X3WetP_est	11.11	34.544

APPENDIX 4

Table S4.1 –

http://www.amjbot.org/content/suppl/2014/04/18/ajb.1300417.DC1/Harbert_AppS1.xls

Table S4.2 –

http://www.amjbot.org/content/suppl/2014/04/18/ajb.1300417.DC1/Harbert_AppS2.doc