Am. J. Trop. Med. Hyg., 70(2), 2004, pp. 105–109

GEOGRAPHIC AND ECOLOGIC DISTRIBUTIONS OF THE *ANOPHELES GAMBIAE* COMPLEX PREDICTED USING A GENETIC ALGORITHM

REBECCA S. LEVINE, A. TOWNSEND PETERSON, AND MARK Q. BENEDICT
Centers for Disease Control and Prevention, Atlanta, Georgia; University of Kansas Natural History Museum, Lawrence, Kansas

Abstract. The distribution of the *Anopheles gambiae* complex of malaria vectors in Africa is uncertain due to under-sampling of vast regions. We use ecologic niche modeling to predict the potential distribution of three members of the complex (*A. gambiae*, *A. arabiensis*, and *A. quadriannulatus*) and demonstrate the statistical significance of the models. Predictions correspond well to previous estimates, but provide detail regarding spatial discontinuities in the distribution of *A. gambiae* s.s. that are consistent with population genetic studies. Our predictions also identify large areas of Africa where the presence of *A. arabiensis* is predicted, but few specimens have been obtained, suggesting under-sampling of the species. Finally, we project models developed from African distribution data for the late 1900s into the past and to South America to determine retrospectively whether the deadly 1929 introduction of *A. gambiae sensu lato* into Brazil was more likely that of *A. gambiae sensu stricto* or *A. arabiensis*.

INTRODUCTION

The *Anopheles gambiae sensu lato* (s.l.) complex contains the world's most efficient vectors of human malaria. Consisting of six named and one unnamed morphologically similar species,[1,2] the complex is primarily responsible for the approximately 80% of global malaria morbidity and mortality that occurs in sub-Saharan Africa.[3] Differences in malaria vector competence among members of the complex have been recognized and are attributed primarily to preferences for feeding on humans versus animals, tendency to enter houses, and ability to recover in numbers after dry seasons.[4] The two members of the complex most responsible for transmission, *A. gambiae* and *A. arabiensis*, are broadly sympatric, although the latter is more broadly distributed in arid regions.[4–6]

An accurate and predictive understanding of the geographic distributions of these species would permit efficient planning of strategies for targeted control or further sampling, detection of competitive interactions, identification of areas in which particular species are potentially involved in transmission, and estimation of risk of introduction to other parts of the world[7] for example from ship cargo transport. Furthermore, intraspecific genetic variation could be associated with specific ecologic niches: e.g., particular *A. gambiae* karyotypes have been correlated with variation in transmission and climate.[8,9]

Factors determining the geographic distributions of *A. gambiae* complex members have been explored via general correlations with climatic factors[6] and using nonlinear equations in combination with spatial mapping tools.[5,10] These studies elucidated climatic factors comprising the species' ecologic niche, but have not demonstrated predictive ability beyond the input data area. Rather, the strong spatial biases in the input data has generally been reflected in the results of previous modeling exercises.[10]

Herein, we apply a machine-learning algorithm to model the ecologic niches occupied by three *A. gambiae* complex species: *A. gambiae* s.s., *A. arabiensis*, and *A. quadriannulatus*, and predict their geographic distributions. The Genetic Algorithm for Rule-set Prediction (GARP[11]) models ecologic niches of species based on relating point-occurrence data to electronic maps of relevant ecological dimensions, producing a heterogeneous set of rules that describe the potential distribution of species in ecological dimensions. The rules defining the niche can then be used to predict the potential distribution of the species elsewhere in time or space by projecting these rules onto appropriate ecologic data.[12] This approach is unique in its ability to construct maps that relate several ecologic factors simultaneously to point occurrence data and in its creation of heterogeneous sets of rules to define the niche; the result is finely resolved distributional predictions at a continental scale.

MATERIALS AND METHODS

Data and statistical test of predictions. An existing *A. gambiae* s.l. dataset[6] was supplemented with additional material from references for a total of 581, 501, and 86 unique occurrence points for *A. gambiae*, *A. arabiensis*, and *A. quadriannulatus* respectively, where an occurrence is equivalent to a georeferenced collection site for a particular species. Occurrence point accuracy was at least to the minute and to the second where greater resolution information was available. Fourteen environmental data layers were considered for inclusion, of which 12 were selected for final models based on error patterns in preliminary analyses. The pixel size was 0.1 degrees. Environmental data layers used in the initial assessments of ecologic niche dimensions were annual mean temperature, annual mean maximum temperature, annual mean minimum temperature, daily temperature range, frost days, topographic aspect, flow accumulation, topographic index, annual mean precipitation, wet days, elevation, and vapor pressure (Intergovernmental Panel on Climate Change, Geneva, Switzerland), tree cover, land-use/land-cover (Department of Geography, University of Maryland, College Park, MD). All except elevation and vapor pressure were used for final models.

We assessed the statistical significance of model predictions using an extrinsic and independent test data set: a randomly selected half of 28 African countries containing sufficient data points was used to build models and to predict species' distributions in the other 14 countries. Data points were stratified into datasets for model building (training data) and model testing (test data) according to an arbitrary criterion: country. We first selected countries from which sufficient sampling points were both available and broadly distributed across the country. Sufficient data points were not available for *A. quadriannulatus* for this purpose. From among these 28

countries, 14 were selected at random for model training or model testing.

Ecologic niche models were created using GARP, a genetic algorithm specifically designed for this challenge,[11] recently made publicly available with a simple user interface (beta.lifemapper.org/desktopgarp). After creating 100 GARP models, a final predicted map was created by summing the five best-subset models.[13] Test data, which were not involved in model building, were overlaid on predicted distributions, and model predictions assessed using a chi-square analysis.[14] The test data from the excluded countries thus served the purpose of independent pseudo-collections to which the predictions of our models were applied (Figure 1A). The best subset of models developed, defined as those models with the lowest training dataset point omission and median commission values,[13] was used to predict species' occurrences in the 14 excluded countries. Chi-square tests were performed to assess the significance of the model's prediction. In all cases, the distributional predictions produced highly significant predictions (Table 1 and Figure 1) even though the data points and environmental characteristics associated with the excluded countries were ignored in model construction.

RESULTS

Final predictions of the distribution of complex members. We derived final predicted distributions using the same environmental layers, but including all point-occurrence data (i.e., without subsetting countries for model testing). The maps predict general sympatry of *A. arabiensis* and *A. gambiae*, with *A. arabiensis* being more widely distributed, par-

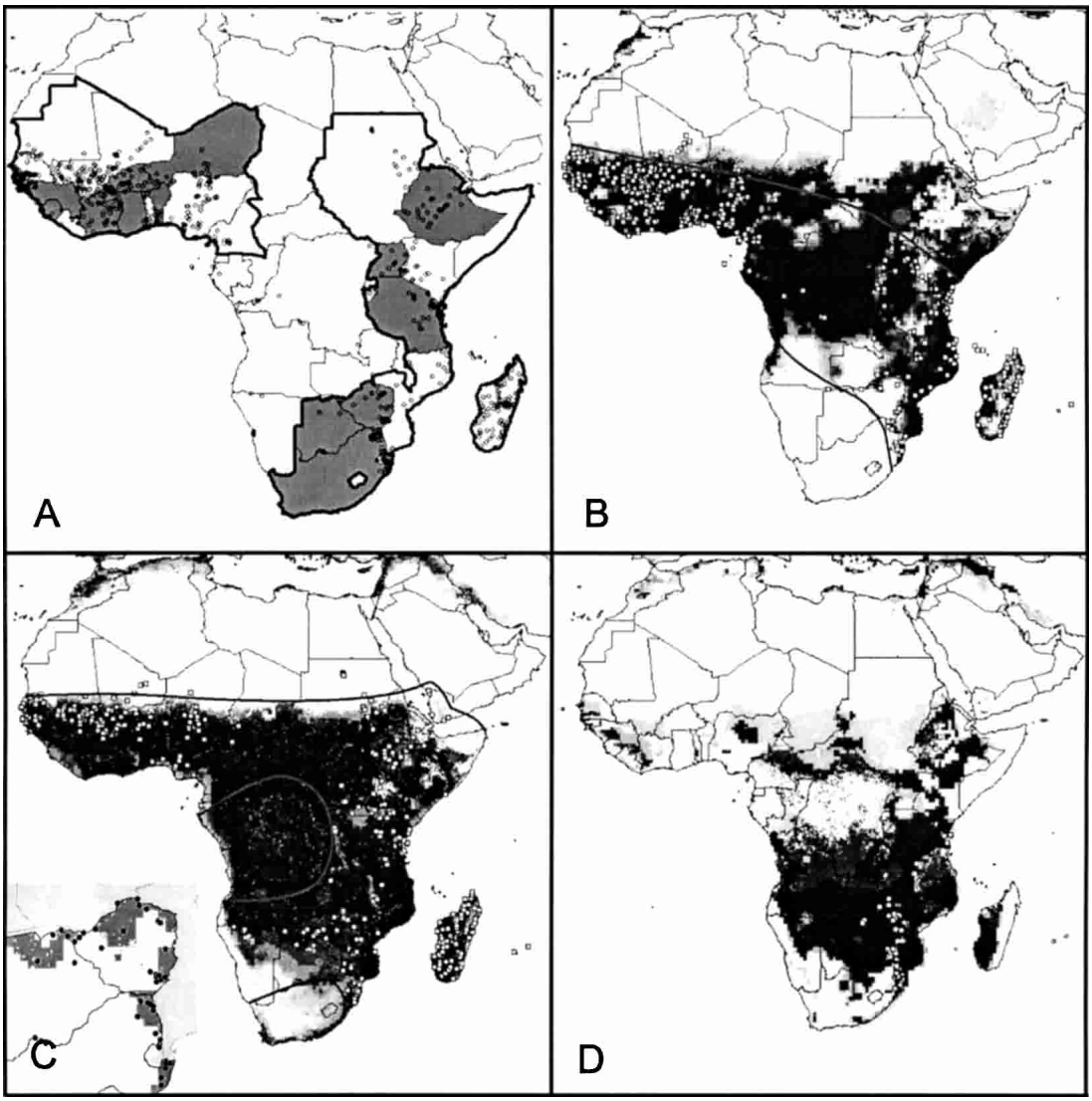

FIGURE 1. Test of the statistical robustness of predictions derived from ecologic niche models for the *Anopheles gambiae* complex and final maps. **A**, Map of Africa showing countries from which distributional data were used to build models (**bold outline**) and those from which distributional data were used to test models (**gray shading**) showing sample points overlaid. The **inset** in 1C shows detail of prediction of the southern range limit of *A. arabiensis* compared with the final prediction. Predicted distributions of **B**, *A. gambiae*, **C**, *A. arabiensis*, and **D**, *A. quadriannulatus*. The **circles** in Ethiopia mark sites from which an unnamed sibling species of *A. quadriannulatus* was identified.[2] The **lines** drawn on the maps for *A. gambiae* and *A. arabiensis* indicate the approximate limits of the distribution of *A. gambiae* and *A. arabiensis* according to the World Health Organization.[15] In all maps, the **darker shading** indicates greater model predictive agreement of presence.

TABLE 1
Statistical tests of distributional predictions for *Anopheles gambiae* complex species*

Species	Test points (no.)	Correctly predicted (no.)	Proportion of area predicted present	Expected points correctly predicted (no.)	Chi-square	*P*
A. arabiensis	162	123	0.542	87.76	79.528	4.75×10^{-19}
A. gambiae	188	172	0.492	83.04	208.066	3.63×10^{-47}

Models were based on occurrence data from 14 African countries, and tests were based on the independent test points from the remaining 14 African countries.

ticularly in south-central Africa (Figure 1B and C). This prediction is generally consistent with that of Lindsay and others,[5] who presented climate suitability zones for these species. Both their maps and ours (Figure 1) differ from the generalized distribution map[15] that illustrates the absence of *A. arabiensis* across most of central Africa (e.g., Angola, Congo, Democratic Republic of Congo, and Gabon). This discrepancy in the central Africa region can be explained in three ways: 1) both our predictions and those of Lindsay and others of suitable habitat for *A. arabiensis* are incorrect, 2) the existing generalized map is incorrect, or 3) the predictions and generalized map correctly identify the potential and realized niches, respectively, but some unidentified biotic or abiotic factor precludes *A. arabiensis* from establishing in that area. Regardless, this area should be sampled intensively to determine conclusively which species occur. Other differences between our maps and the World Health Organization distributions are seen in our predictions of suitable environments for *A. gambiae* in northeastern Africa, in the lower elevations of Ethiopia, northern Kenya, southern coastal Somalia, and southeastern Sudan.

Previously, predictions of the geographic distributions of these species have been developed based on a similar data set, but using very different analytical approaches.[10] The results of the two modeling efforts contrast sharply, particularly in central Africa, where sampling was most sparse (Figure 1A): our models predicted broad areas of presence for *A. gambiae* and *A. arabiensis*, whereas the previous models did not. We suggest that this difference results from the inability of the previous, regression-based inferences to predict into broad unsampled areas. For example, the regression-based predictions for *A. gambiae* assigned the same probability of presence to south-central Africa as the inhospitable Sahara Desert. Also, comparison with historical maps shows large areas of known occurrence of *A. arabiensis* in coastal west Africa not predicted by the regression-based models. Apart from this dissatisfaction with the result, we note their lack of independent validation of model predictions, uneven interpretation of kappa statistics, and over-reliance on assumptions of absence based on data not well suited to establish absences with confidence because 96% of the samples in the database of Coetzee and others[6] consist of 10 individuals or less.

DISCUSSION

In all of our models, two spatial discontinuities appear that coincide with observed reductions in gene flow between populations. We predict extensive areas of relative niche unsuitability in the eastern Rift Valley *A. gambiae* complex in Kenya.[16] Genetic breaks have also been recognized within *A. quadriannulatus*, given the discovery of a sibling species in Ethiopia.[2] Again, our models indicate a corresponding geographic disjunction in the habitable range of *A. quadriannulatus* separating southern and northern populations. Indeed, although the sites from which the sibling species was identified are surrounded by habitats suitable for *A. quadriannulatus*, the distributional area of the sibling species is not predicted to be habitable. This result thus suggests that the ecologic niches of the two *A. quadriannulatus* forms are distinctly different. No such potential barriers to gene flow are apparent among mainland *A. arabiensis* populations.

The relatively large number of occurrence points for *A. gambiae* and *A. arabiensis* makes possible generalizations about differences between their ecologic niches. Our analysis generally agrees with previous conclusions that relative to *A. arabiensis*, *A. gambiae* inhabits wetter and warmer environments (26.7°C versus 24.6°C annual mean temperature; 28 cm versus 22 cm annual mean precipitation; 0.87 versus 2.57 frost days annually). Much greater detail of visualization of ecologic niches of these forms is possible, but is not presented herein for reasons of space.

The influence of each environmental data layer on model predictions was assessed via a jackknifing procedure, which identifies layers whose exclusion most influenced predictions (Figure 2).[13] Each of the 13 data layers was eliminated from analyses sequentially, and 10 models developed using GARP iterations. These maps were summed to produce a composite map, each pixel of which had a value between 0 and 10, representing the number of replicate models predicting presence. These maps were compared pixel-by-pixel with a similar map created using all layers. Agreement of maps was calculated as weighted kappa values[17] based on pixels in agreement. Frost days influenced predictions for *A. gambiae* particularly strongly. In contrast, both climatic and topographic characteristics influenced *A. quadriannulatus* distribution, and no environmental variables uniquely affected *A. arabiensis* models markedly.

We used these models to test the likely role of different complex member species in a historical event: the establishment of *A. gambiae* s.l. in northeastern Brazil during the1930s via accidental introduction, probably from Senegal. This introduction and subsequent spread resulted in tens of thousands of deaths from epidemic malaria. These epidemics ended after a military style eradication program led by Fred Soper.[18] Because *A. gambiae* was yet not recognized as a complex of several species, the identity of the member of the complex that was responsible remains a mystery. White speculated that it was *A. arabiensis*, given its probable origin in Senegal;[4] however, three *gambiae* complex species occur in that country: *A. gambiae*, *A. arabiensis*, and *A. melas.*[19] Although present in Senegal, *A. melas* is not a major vector species of human disease and is believed to breed in brackish water for which appropriate data layers were unavailable. Thus, using climate data from 1960 to 1990 to build African

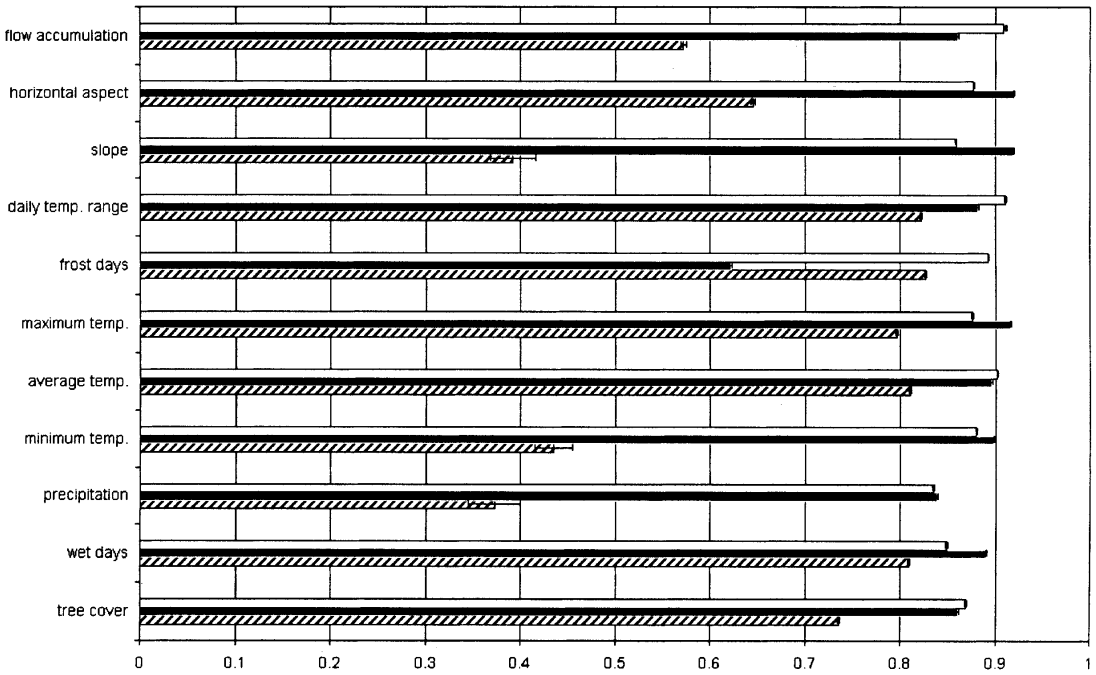

FIGURE 2. Kappa values calculated from jackknife experiments. Error bars indicate 95% confidence intervals. Layers excluded from model building are listed on the vertical axis. The open, filled, and hatched bars represent values for *Anopheles arabiensis*, *A. gambiae*, and *A. quadriannulatus*, respectively. temp. = temperature.

native-range models, we projected the niches of the two primary vectors among these species to the Americas for 1930–1960 climate regimes. These models indicate that *A. arabiensis* would have had an extremely limited potential distribution (Figure 3A), whereas *A. gambiae* had widespread suitable habitat near the point of introduction and broadly throughout the Americas and Caribbean (Figure 3B). These results suggest that Soper's heroic eradication campaign stopped the spread of this vector at the doorstep of extensive areas of favorable niche, in which eradication would have been impossible.

Discrepancies between our predicted distributions and the believed realized distribution may occur for several reasons unrelated to either the modeling method or the data analyzed: local elimination, failure to overcome geographic and climatic barriers to introduction, and biotic interactions that were not considered in the model. While the distribution of species with which a species may interact may be known, the effect of the interaction often is not. Therefore, we were unable to consider these possible factors in modeling the mosquitoes we studied. Moreover, our models did not exclude data points that may have occupied relatively small portions

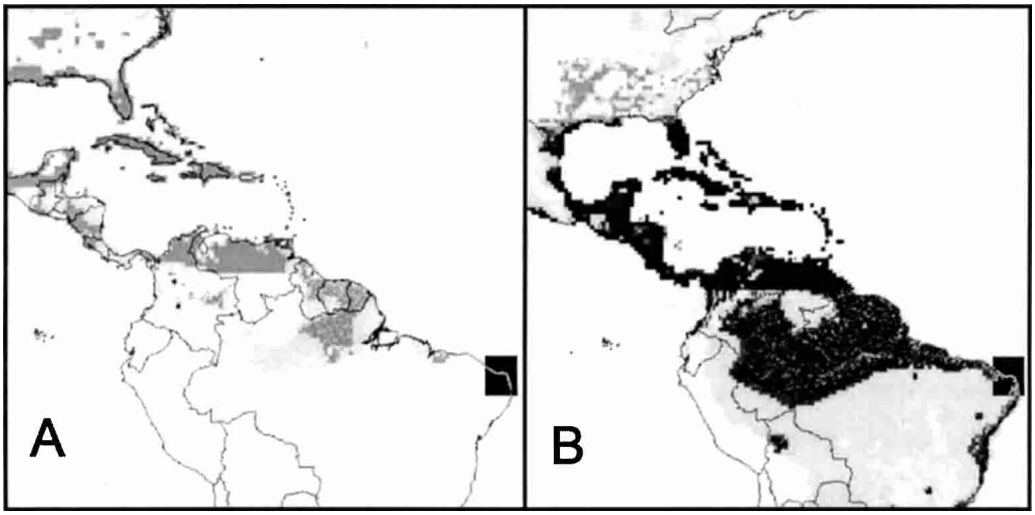

FIGURE 3. Predicted ranges of **A**, *Anopheles arabiensis* and **B**, *A. gambiae* developed from the native range models (Figure 1) and projected onto climate data from 1931 to 1960 for South America. In all maps, **darker shading** indicates greater model agreement in prediction of presence. The area in which *A. gambiae* s.l. became established and was eradicated is indicated by the **black square** in northeastern Brazil.

of the pixel within which they fall; such error would represent a pixel as suitable when in reality, only a small portion of it would be. Easily identifiable examples that are often available in spatial databases and which could be considered are rivers, streams, and small standing bodies of water that are subject to flooding. Including such exceptional data in the analysis results in over-prediction of potential distribution when modeling species such as mosquitoes that have restricted flight ranges and do not migrate.

The implications of these new methodologies are numerous. Species' distributions can be inferred and predicted with high precision, permitting extrapolation of known information to a much broader area and anticipation of distributional patterns that would be otherwise poorly understood.[12,14,20] Species' native distributions can be used to infer ecologic and distributional potential in other regions as an invasive species.[7] Changes to be expected in species' geographic distributions that may result from ongoing global climate change can be predicted.[21] Implications for potential bioterrorism applications are also clear, particularly when disease organisms are involved.[20] In summary, ecologic niche modeling offers a powerful tool for understanding distributional phenomena related to biodiversity: in the present case, offering a solution to the 70-year-old mystery of which *Anopheles* species caused one of the most serious malaria outbreaks in the history of the New World.

Received April 14, 2003. Accepted for publication July 5, 2003.

Financial support: This research was supported in part by an appointment of Rebecca S. Levine to the Emerging Infectious Diseases (EID) Fellowship Program, administered by the Association of Public Health Laboratories (APHL) and funded by the Centers for Disease Control and Prevention (CDC).

Authors' addresses: Rebecca S. Levine, Entomology Branch, Division of Parasitic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Mailstop F-22, 4770 Buford Highway NE, Atlanta, GA 30341-3717, Telephone: 770-488-7318, Fax: 770-488-4258, E-mail: Rlevine@cdc.gov. A. Townsend Peterson, Natural History Museum and Biodiversity Research Center, University of Kansas, Lawrence, KS 6604, Telephone: 785-864-3926, E-mail: mexbidiv@lark.cc.ukans.edu. Mark Q. Benedict, Entomology Branch, Division of Parasitic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Mailstop F-22, 4770 Buford Highway NE, Atlanta, GA 30341-3717, Telephone: 770-488-4987, Fax: 770-488-4258, E-mail: Mbenedict@cdc.gov.

Reprint requests: Mark Q. Benedict, Entomology Branch, Division of Parasitic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Mailstop F-22, 4770 Buford Highway NE, Atlanta, GA 30341-3717.

REFERENCES

1. Davidson G, Paterson HE, Coluzzi M, Mason GF, Micks DW, 1967. The *Anopheles gambiae* Complex. Wright JW, Pal R, eds. *Genetics of Insect Vectors of Disease.* Amsterdam: Elsevier, 211–250.
2. Hunt RH, Coetzee M, Fettene M, 1998. The *Anopheles gambiae* complex: a new species from Ethiopia. *Trans R Soc Trop Med Hyg 92:* 231–235.
3. Breman JG, Egan A, Keusch GT, 2001. The intolerable burden of malaria: a new look at the numbers. *Am J Trop Med Hyg 64 (suppl):* iv–vii.
4. White GB, 1974. *Anopheles gambiae* complex and disease transmission in Africa. *Trans R Soc Trop Med Hyg 68:* 278–298.
5. Lindsay SW, Parson L, Thomas CJ, 1998. Mapping the ranges and relative abundance of the two principal African malaria vectors, *Anopheles gambiae sensu stricto* and *An. arabiensis*, using climate data. *Proc R Soc Lond B Biol Sci 265:* 847–854.
6. Coetzee M, Craig M, le Sueur D, 2000. Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. *Parasitol Today 16:* 74–77.
7. Peterson AT, Vieglais DA, 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *BioScience 51:* 363–371.
8. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA, 1979. Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. *Trans R Soc Trop Med H 73:* 483–497.
9. Toure YT, Petrarca V, Traore SF, Coulibaly A, Maiga HM, Sankare O, Sow M, Di Deco MA, Coluzzi M, 1998. The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia 40:* 477–511.
10. Rogers DJ, Randolph SE, Snow RW, Hay SI, 2002. Satellite imagery in the study and forecast of malaria. *Nature 415:* 710–715.
11. Stockwell DRB, Peters D, 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *Int J Geogr Inf Sci 13:* 143–158.
12. Peterson AT, Ball LG, Cohoon KP, 2002. Predicting distributions of Mexican birds using ecological niche modelling methods. *Ibis 144:* E27–E32.
13. Anderson RP, Lew D, Peterson AT, 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecol Model 162:* 211–232.
14. Peterson AT, 2001. Predicting species' geographic distributions based on ecological niche modeling. *The Condor 103:* 599–605.
15. White GB, 1989. *Malaria. Geographical Distribution of Arthropod-Borne Diseases and their Principal Vectors*. Geneva: World Health Organization, WHO/VBC, 7–22.
16. Lehmann T, Hawley WA, Grebert H, Danga M, Atieli F, Collins FH, 1999. The Rift Valley complex as a barrier to gene flow for *Anopheles gambiae* in Kenya. *J Hered 90:* 613–621.
17. *SAS OnlineDoc, (rtm),* 2002. Cary, NC: SAS Institute, Inc.
18. Soper FL, Wilson DB, 1943. *Anopheles gambiae* in Brazil: 1930 to 1940. New York: The Rockefeller Foundation.
19. Petrarca V, Vercruysse J, Coluzzi M, 1987. Observations on t he *Anopheles gambiae* complex in the Senegal River Basin. *West Afr Med Vet Entomol 1:* 303–312.
20. Peterson AT, Sanchez-Cordero V, Beard CB, Ramsey JM, 2002. Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerg Infect Dis 8:* 662–667.
21. Peterson AT, Ortega-Huerta MA, Bartley J, Sanchez-Cordero V, Soberon J, Buddemeier RH, Stockwell DR, 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature 416:* 626–629.