PNAS
www.pnas.org

Supplementary Information for

A machine learning approach to map landscape connectivity in *Aedes aegypti* using genetic and environmental data

Evlyn Pless, Norah P. Saarman, Jeffrey R. Powell, Adalgisa Caccone, Giuseppe Amatulli

Evlyn Pless and Giuseppe Amatulli
Email: espless@ucdavis.edu and giuseppe.amatulli@yale.edu

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Supplementary Information Text

Details on the Genetic Analysis: Genetic drift

In this supplemental material, we include tests for the effect of 7 years (~70 *Aedes aegypti* generations) of genetic drift in two ways: a) simulating expected differentiation due to 70 generations of drift, and b) using empirical data to calculate genetic differentiation for samples collected from the same population sampled at different time-points.

We ran simulations using the individual-based forward-time program simuPOP [1, 2]. We simulated a population with 12 loci, random mating, no migration, and an N_e equal to 500. We took random samples of 20 individuals from generation 0 (after 100 generations of burn-in) and from generation 70, and used Genepop [3] to calculate the CSE between the two samples. This process was repeated three times, and the resulting CSE values were 0.173, 0.236, and 0.168. These values are similar to the temporal empirical results and significantly less than the mean CSE results between different sites reported in the submitted paper (p=0.02), giving us additional confidence that genetic drift is not confounding the model. The three replicates produced F_{ST} values of 0.173, 0.236, and 0.168, with corresponding p-values of 0.13, 0.00000054, and 0.10. Although these are not statistically different than values in the paper (p = 0.13), two of three are not statistically significant from zero.

Additionally, we identified 12 populations in North America (including 5 that are included in the submitted paper) for which we have samples from 2, 3, or 4 different years. In total this yields 35 time interval pairs, and these time intervals range from 1 to 12 years (mean = 3.4 years). The mean CSE among these points is 0.22 ± 0.07, significantly lower than the 0.34 ± 0.065 mean reported for CSE values in the manuscript ($p < 10^{-11}$). (Just considering the populations that also appear in the manuscript, the mean time interval is 3.3 years and the CSE is 0.21 ± 0.70.) A linear regression shows no correlation between the length of the time interval and the value of CSE (adjusted R^2 = -0.0037, p-value = 0.36) (Fig. 15A). Similarly, the mean linearized F_{ST} (0.051 ± 0.087) for these time intervals. The temporal samples were significantly lower than the linearized FST values presented in the paper (0.086 ± 0.043) (p=0.017). A linear regression shows no correlation between the length of the time interval and the value of linearized F_{ST} (adjusted R^2 = -0.0065, p-value = 0.38) (Fig. 15B). These results indicate the genetic distance caused by resampling the same site at different years can be explained by a small amount of noise, possibly related to sampling error. Genetic drift is evidently not playing a large role, even for time samples taken more then 70 generations apart.

Details on the modeling process:

Leave-two-out cross-validation

A concern with the leave-one-out cross-validation (LOOCV) is that we would expect the error values of the training dataset and the full dataset to converge as the size of the training dataset increases. To ensure that the root mean square error (RMSE) of the LOOCV is not simply due to the large training dataset, we also ran a leave-two-out cross-validation (LTOCV) using CSE, in which two points and all their affiliated pairs were withheld as the testing dataset for each of the 16 runs (38 points/2). While the LOOCV testing datasets only contain 5.2% of the data (37/706 pairs of points), the LTOCV testing datasets each contain 11.9% of the data (75/706 pairs of points), very similar to the proportion of data withheld for testing during the widely-used ten-fold cross-validation procedure.

Comparing LOOCV and LTOCV we found that the mean $RMSE_{test}$ and $RMSE_{train}$ values showed essentially no change (Table S9). Additionally, the mean $RMSE_{train}$ for the cross-validations (0.036) is similar to $RMSE_{full}$ for the full model run using CSE (0.035). These results strongly suggest that the consistent values RMSE between the cross-validations and full model

run are not simply due to the large size of the training dataset, but rather to the model's performance.

Linear regression

As a basis for comparison, we also fit our model using a standard linear regression in place of Random Forest (RF). In order to highlight certain advantages of RF, we kept the input data the same in the linear regression model as the RF model, including using all 29 environmental and anthropogenic spatial datasets. We used CSE as genetic distance, and we used all the genetic data to build the model (i.e. no cross-validation). We modeled genetic connectivity using straight lines (iteration 0) and one round of least cost path analysis (iteration 1), as this was sufficient to demonstrate the issues with this approach.

Modeling genetic distance with straight lines (iteration 0), the R^2 of the linear regression model was 0.433 ($p < 10^{-16}$), much lower than the RF model from the same iteration ($R^2 = 0.618$). The most important variables from the linear regression model ($p<0.001$) were altitude, potential evapotranspiration, precipitation of the wettest month, precipitation of the driest month, and the kernel density map. The prediction surface from the linear regression contained extremely large values (>470,000) that are not within the expected range for the inverse of CSE (2-6) (Fig. S16A). These values likely distort the least cost path analysis. When these high-value outliers were removed, the prediction surface could be visualized, and it showed little spatial detail (Fig. S16A).

Modeling genetic distance with least cost paths (iteration 1), the linear regression model's R^2 was 0.402 ($p < 10^{-16}$), lower than the R^2 from iteration 0 of the same model and much lower than the RF model from the same iteration ($R^2 =0.681$). The most important variables ($p<0.001$) were aridity, human density, friction, and potential evapotranspiration, all different from the most important variables in the first iteration. Again, the prediction surface contained outliers (>55,000,000), and when the outliers were removed the prediction surface showed little spatial detail (Fig. S16B). Although this is a toy model it clearly illustrates some of the advantages of RF over a standard linear regression when modeling complex relationships among many variables, some correlated. The RF approach we employ provides greater accuracy across the distribution of the species, more spatial detail, fewer unreasonable (extreme outlier) predictions of connectivity, and more stable assessment of variable importance.

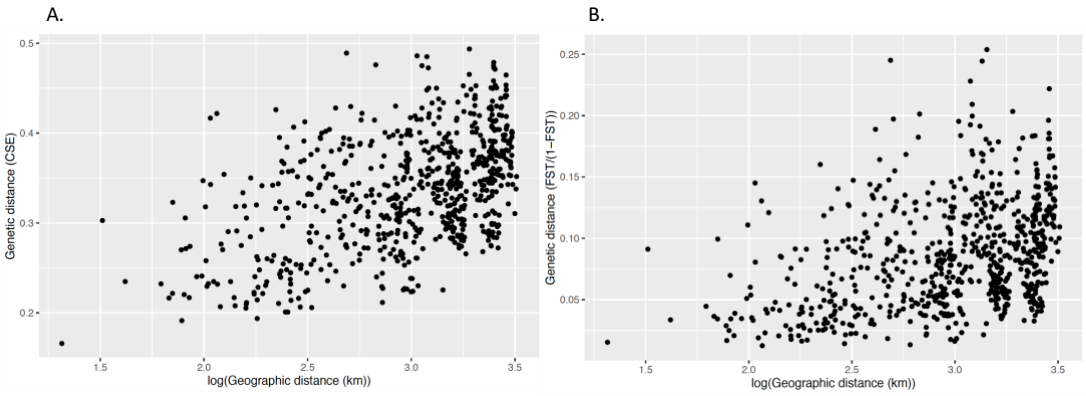


Fig. S1. Relationship between geographic distance and genetic distance for all sites. (A) Log geographic distance vs. CSE and (B) Log geographic distance vs. linearized F_{ST}.

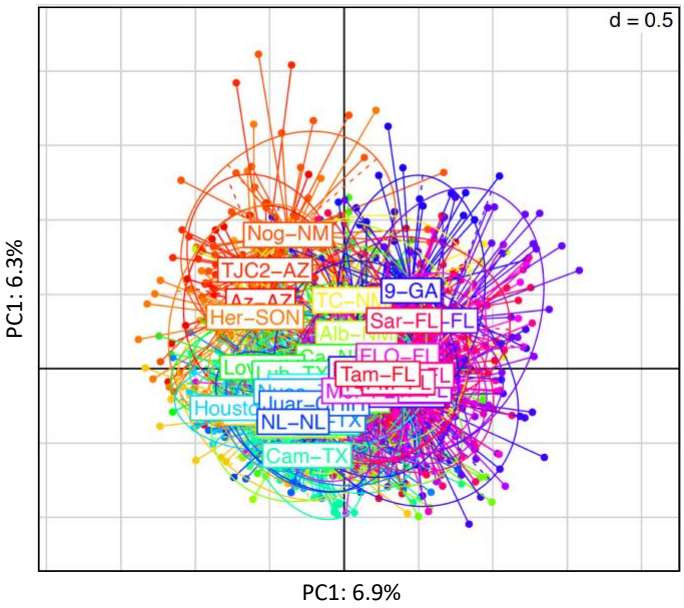


Fig S2. Genetic structure of North America using Principal Component Analysis. Ellipses indicate the distribution of individuals within each population. Populations groups are labeled by their site abbreviation and state. (See Table S1 for full list of sites and corresponding abbreviations.)

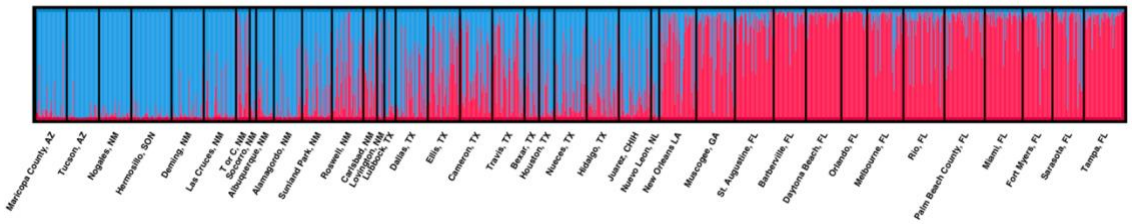


Fig. S3. STRUCTURE plot of North America dataset. Each vertical bar represents an individual, and the proportion of each color assigned to each individual represents the proportion of the individual's ancestry attributable to each of the K theoretical genetic clusters (K=2).

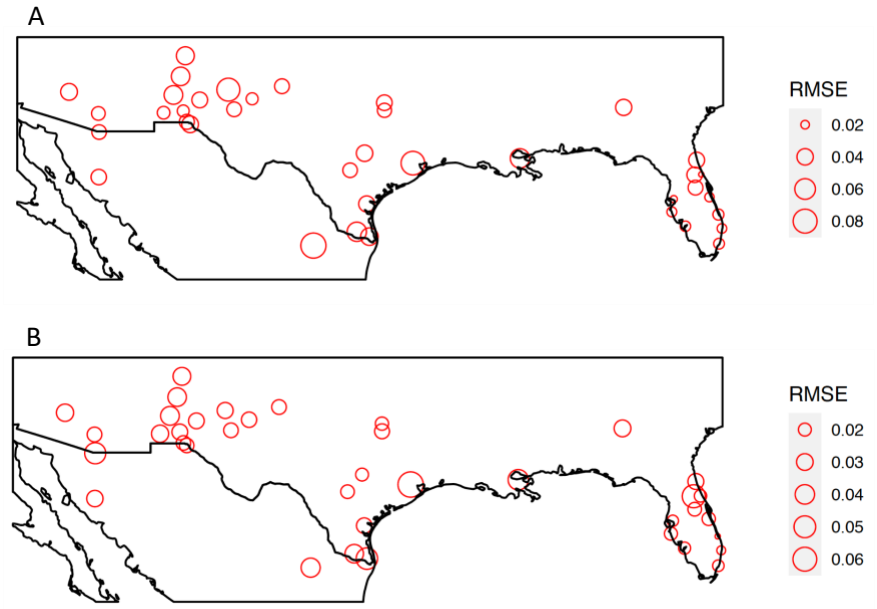


Fig. S4. Root mean square error ($RMSE_{test}$) for each site in the leave-one-out cross-validation for CSE (A) and linearized F_{ST} (B). Circle size corresponds to $RMSE_{test}$ value. In the same vein of Fig. S9, the goal of this model is to determine the influence of spatial autocorrelation on the model. Although there are some clusters of low/high $RMSE_{test}$ values, there are a range of $RMSE_{test}$ values across the map and between points that are in low or highly sampled areas.

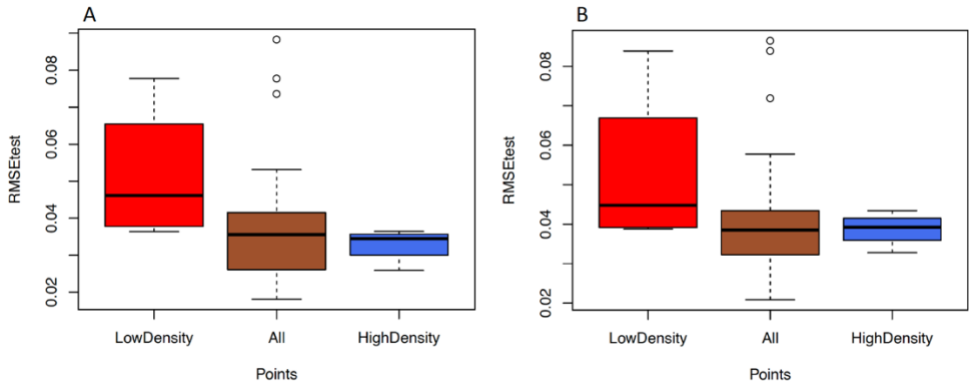


Fig. S5. CSE leave-one-out cross-validation root mean square error ($RMSE_{test}$) before (A) and after (B) weighting the RF bootstrapping. The RF bootstrapping was weighted by the inverse of the kernel density of the lower kernel density site for each pair of sites, ensuring that low density sites were sampled more frequently. The first purpose of this figure is to show that weighting the RF bootstrapping decreases the difference in $RMSE_{test}$ between the high and low density sites. The second purpose is to show the difference in $RMSE_{test}$ between the points with the highest (10%) values on the kernel density map, those with the lowest (10%), and all points. Although the ranges are overlapping, the low density points category has higher and more variable $RMSE_{test}$ than the high density category.

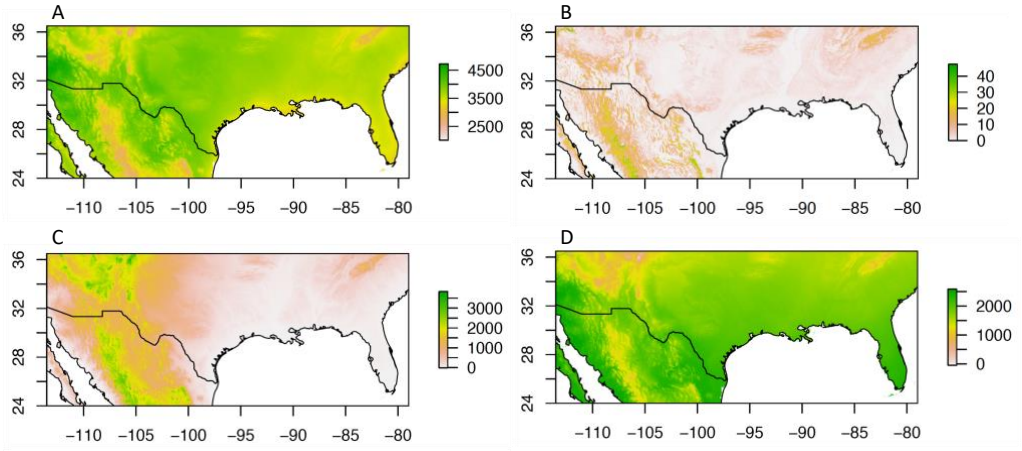


Fig. S6. The four most important variables for the leave-one-out cross-validation using CSE as genetic distance. A. maximum temperature (Celsius x 100), B. slope (degree), C. altitude (meters), and D. mean temperature (Celsius x 100).

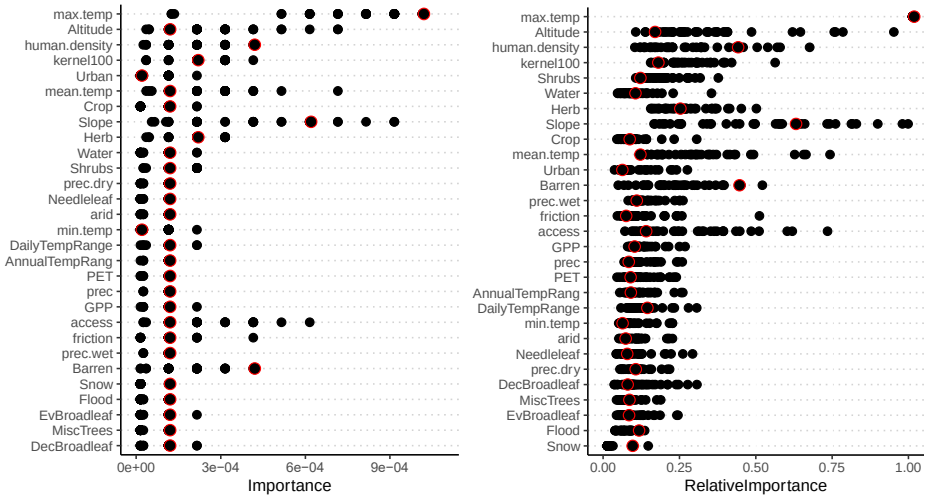


Fig. S7. Importance (left) and relative importance (right) of all variables for leave-one-out cross-validation with CSE as genetic distance. Importance is mean decrease in model accuracy when removing each variable, and relative importance is scaled such that the most important variable has importance equal to 1. Results across all 38 folds are depicted to show the relatively high consistency for which variables were ranked as most or least important. The point circled in red shows the result from the full dataset run for comparison.

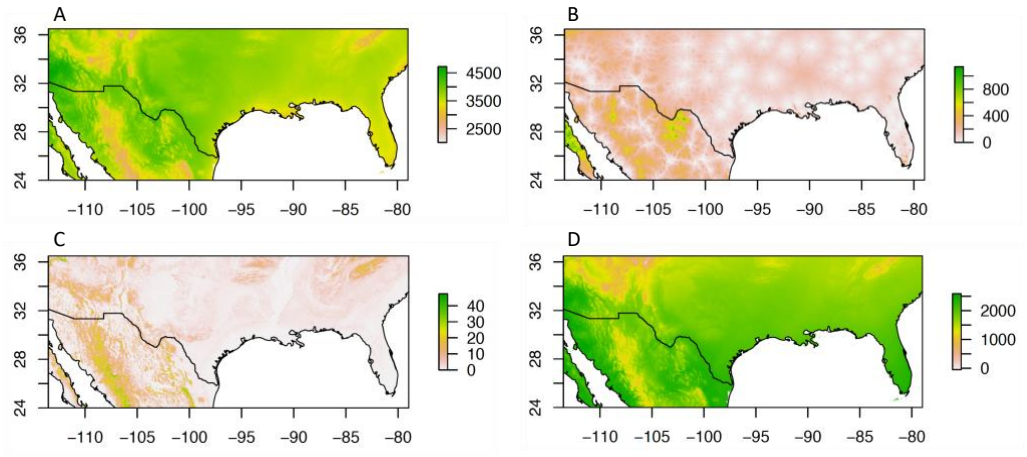


Fig. S8. The four most important variables for the leave-one-out cross-validation using linearized F_{ST} as genetic distance. A. maximum temperature (Celsius x 100), B. accessibility (travel time to the nearest major city), C. slope (degree inclination), and D. mean temperature (Celsius x 100).

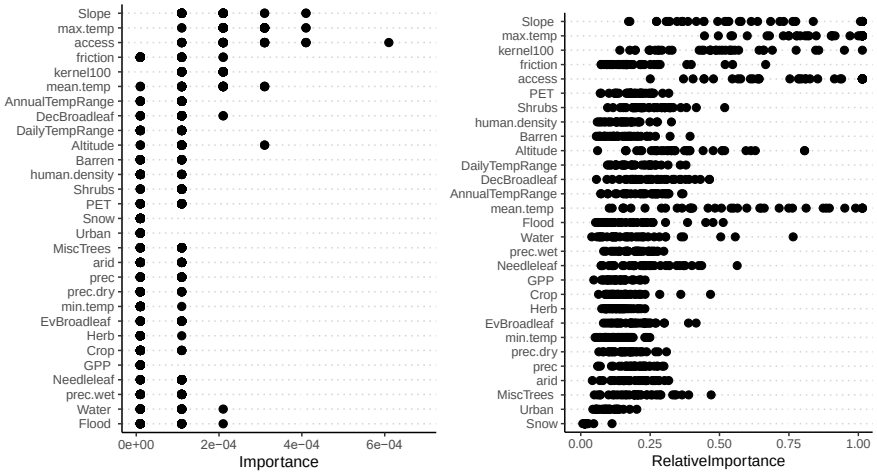


Fig. S9. Importance (left) and relative importance (right) of all variables for leave-one-out cross-validation with linearized F_{ST} as genetic distance. Importance is mean decrease in model accuracy when removing each variable, and relative importance is scaled such that the most important variable has importance equal to 1. Results across all 38 folds are depicted to show the relatively high consistency for which variables were ranked as most or least important.

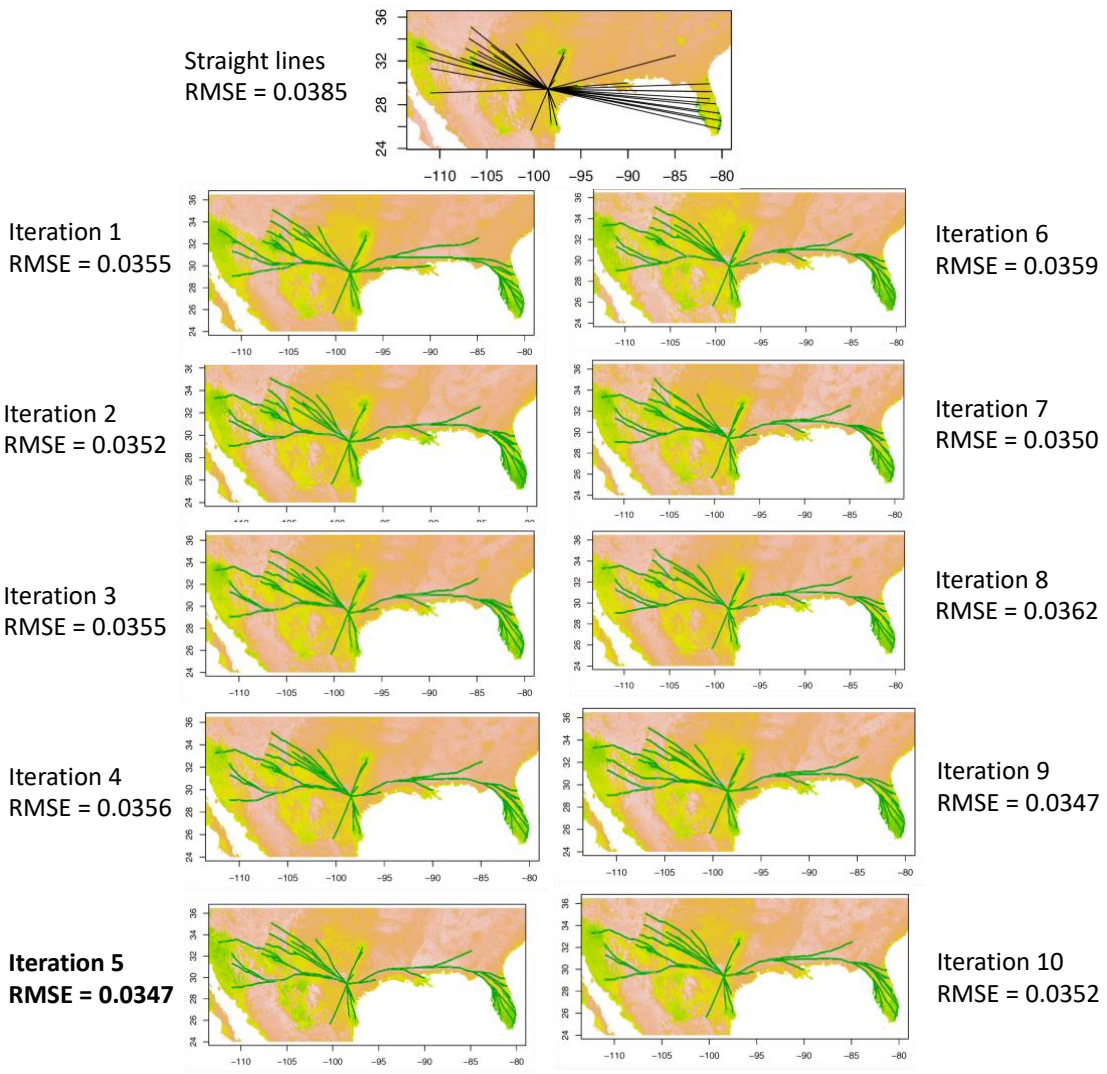


Fig. S10. Straight lines (top row) and least cost path lines using Bexar, Texas as the focal point. We show iterations 1-10 for a full dataset model run using CSE as genetic distance. Behind the lines are the predicted connectivity surfaces generated from the model built using those lines. Each map is labeled with the iteration number and root mean square error of the associated RF model ($RMSE_{full}$). The optimized model was reached after five iterations (the iteration with the lowest $RMSE_{full}$) in this case. In the connectivity surfaces, green is high connectivity and red is low connectivity. Least cost path lines are show in dark green, and the straight lines used to initialize the model are shown in black.

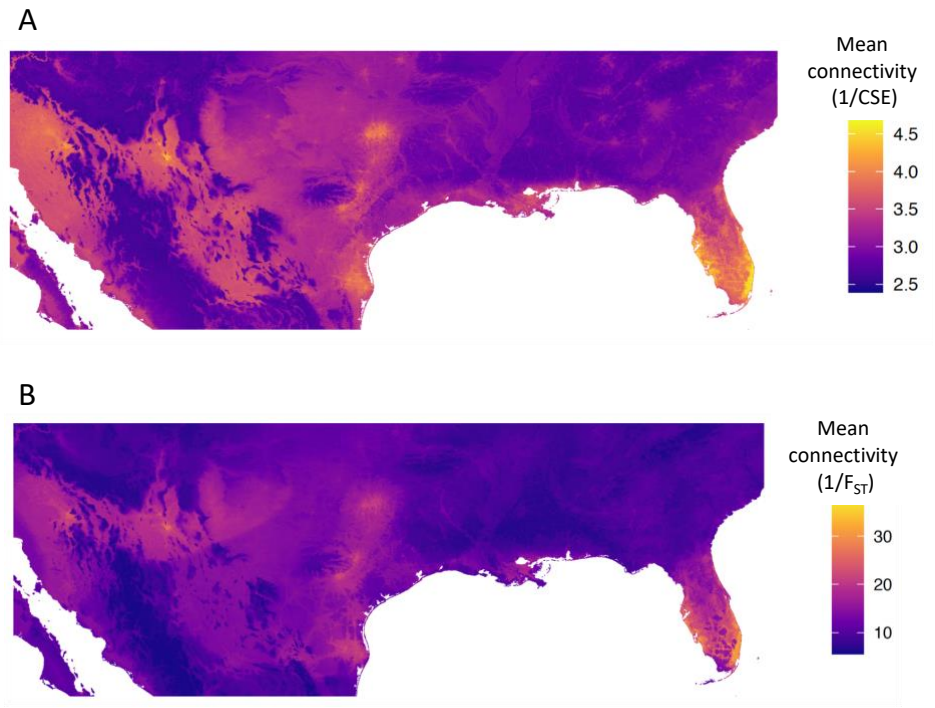


Fig. S11. Mean of the 38 leave-one-out cross-validation optimized resistance surfaces for (A) CSE and (B) linearized F_{ST}. Light colors (yellow) indicate high connectivity, while dark colors (purple) indicate low connectivity.

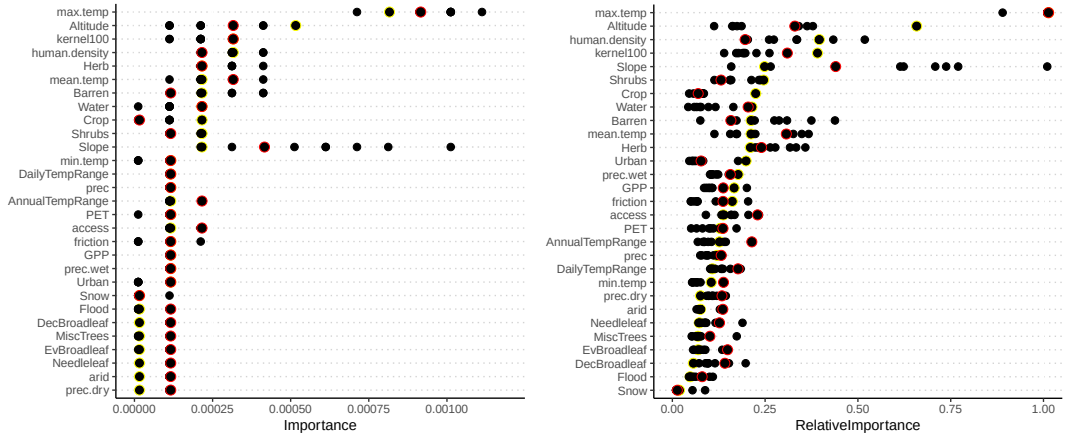


Fig. S12. Importance (left) and relative importance (right) of all variables for the full dataset model using CSE as genetic distance. Importance is mean decrease in model accuracy when removing each variable, and relative importance is scaled such that the most important variable has importance equal to 1. The results are shown for all 10 iterations to show the relatively high stability for which variables are chosen as most/least important. The point circled in red represents the result from the best iteration (lowest $RMSE_{test}$), and the point in yellow is the result from the initialization of the model with straight lines (iteration 0).

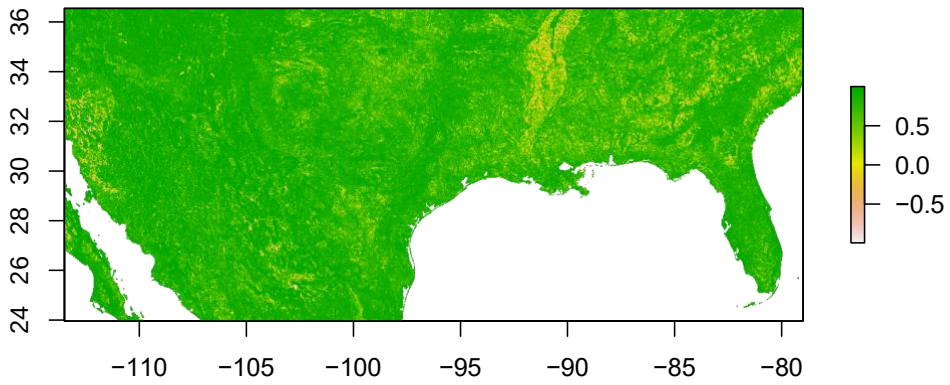


Fig. S13. Pearson correlation between the mean resistance map generated by the leave-one-out cross-validation and the resistance map generated by the full dataset run, using CSE as genetic distance in both cases. Darker green show areas of high correlation, while yellow shows areas of low correlation.

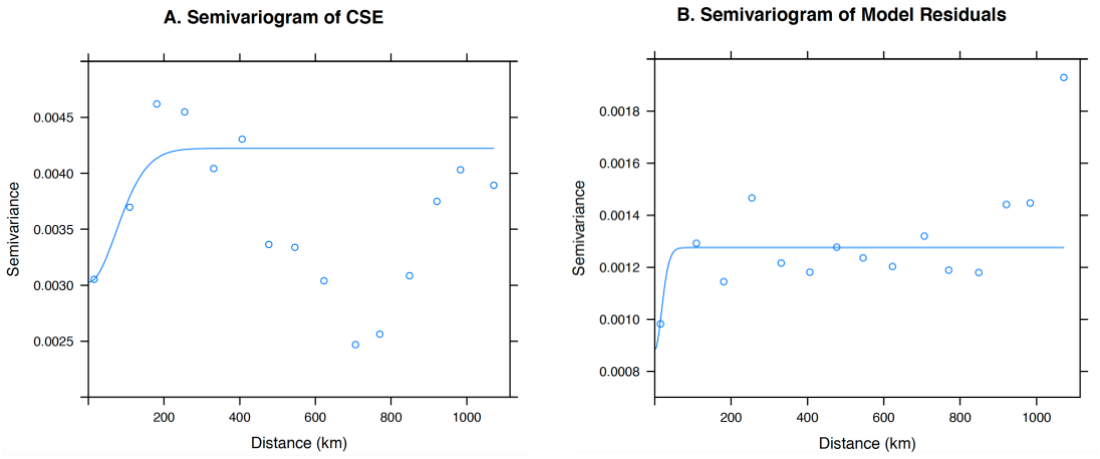
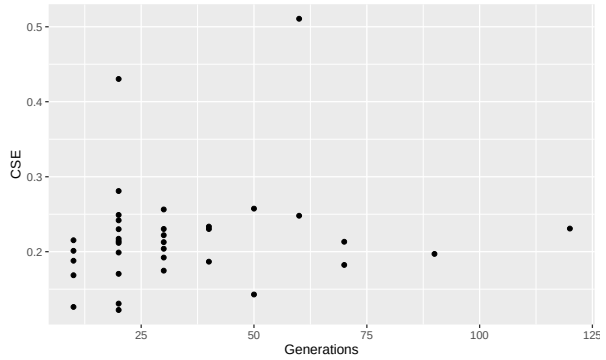


Fig. S14. Semivariograms to show the influence of spatial autocorrelation, i.e. systematic spatial variation in a variable. The x axis is distance bins, the y axis is semivariance, and the blue line shows the best model fit. Spatial autocorrelation and geographic distance influence CSE up until 200km, as shown by increasing semivariance up until this distance (A). There is a large reduction of the impact of spatial autocorrelation on the semivariance of the model residuals (observed – predicted CSE), as shown the leveling of the model fit line at a much shorter distance (B).

A.



B.

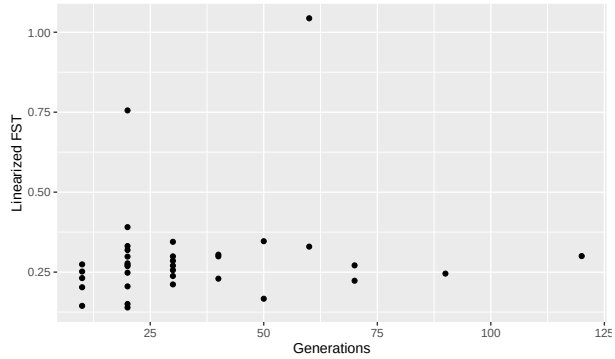
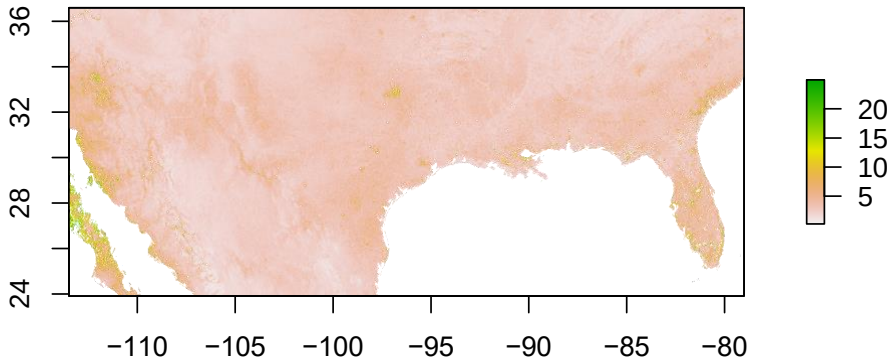


Fig. S15. The effect of interval time (assuming 10 generations/year) and genetic distance for populations sampled in multiple years. A. CSE, B. Linearized F_{ST}.

A.



B.

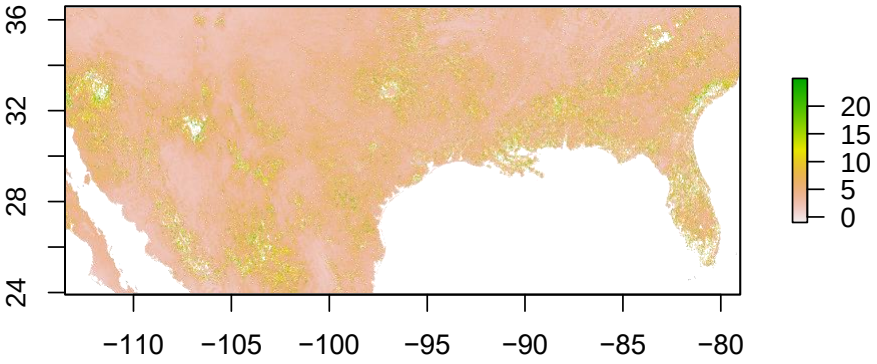


Fig. S16. Predicted connectivity surfaces for full model with CSE and using a standard linear regression in place of Random Forest, for iteration 0 (A) and iteration 1 (B).

Table S1. Sampled locations, corresponding abbreviation, latitude, longitude, sampling year, number of individuals sampled, and whether the data are being published here for the first time.

Site location	Short	Latitude	Longitude	Year	N	New
Maricopa County, AZ, USA	Az	33.2918	-112.4291	2013	39	no
Tucson, AZ, USA	TJC2	32.2226	-110.9747	2012	40	no
Nogales, NM, USA	Nog	31.3012	-110.9381	2013	40	no
Hermosillo, SON, MEX	Her	29.073	-110.9559	2013	50	no
Deming, NM, USA	Dem	32.2593	-107.7401	2017	40	yes
Las Cruces, NM, USA	LC18	32.34568	-106.7661	2018	40	yes
Truth or Consequences, NM	TC	33.137	-107.2526	2017	17	yes
Socorro, NM, USA	Soc	34.0578	-106.8976	2017	8	yes
Albuquerque, NM, USA	Alb	35.0788	-106.6627	2018	22	yes
Alamagordo, NM, USA	Ala	32.8909	-105.9484	2017	35	yes
Sunland Park, NM, USA	SP	31.8073	-106.587	2017	37	yes
Roswell, NM, USA	Ros	33.4009	-104.5294	2017	39	yes
Carlsbad, NM, USA	Car	32.427	-104.243	2017	17	yes
Lovington, NM, USA	Lov	32.9476	-103.3504	2017	9	yes
Lubbock, TX, USA	Lub	33.5699	-101.8727	2017	14	yes
Dallas, TX, USA	Dall	32.75426	-96.79689	2017	40	no
Ellis, TX, USA	El	32.3782	-96.793	2017	40	yes
Cameron, TX, USA	Cam	26.125309	-97.53918	2015	40	no
Travis, TX, USA	Tr	30.2526	-97.778	2017	40	yes
Bexar, TX, USA	Bex	29.4112	-98.4954	2017	18	yes
Houston, TX, USA	Houston	29.7628	-95.3831	2011	19	no
Nueces, TX, USA	Nuec	27.7439	-97.672	2017	40	yes
Hidalgo, TX, USA	Hid	26.3651	-98.1638	2017	40	yes
Juarez, CHIH, MEX	Juar	31.683	-106.4249	2017	40	yes
Nuevo Leon, NL, MEX	NL	25.6803	-100.3133	2017	10	yes
New Orleans LA, USA	NO	29.9984	-90.07611	2012	46	no
Muscogee, GA, USA	18 and 9	32.5223	-84.9341	2011	48	no
St. Augustine, FL, USA	StA	29.9134	-81.31774	2017	48	no
Barberville, FL, USA	Bb	29.1861	-81.4209	2017	40	no
Daytona Beach, FL, USA	DB	29.2034	-81.0884	2017	44	no
Orlando, FL, USA	Orl	28.5496	-81.3754	2014	32	no
Melbourne, FL, USA	Mel	28.0785	-80.6838	2014	45	no
Rio, FL, USA	FLO	27.2183	-80.24	2014	51	no
Palm Beach County, FL, USA	PBC_filter3	26.53	-80.0658	2013	50	no
Miami, FL, USA	Miami	25.7753	-80.2089	2011	47	no
Fort Myers, FL, USA	FM	26.6398	-81.8745	2014	37	no
Sarasota, FL, USA	Sar	27.3509	-82.5484	2014	39	no
Tampa, FL, USA	Tam	27.9816	-82.4526	2014	50	no

Table S2. Spatial data list, sources, resolution of the original dataset, and resampling method used (if any).

Abbreviation	Definition	Source name	Original resolution
arid	Global Aridity Index	CGIAR CSI	1km^2
access	Travel time to the nearest city of 50,000 inhabitants	Weiss *et al.* 2018	~1km^2, bilinear resampling
prec	Annual precipitation	CHELSA climate data	1km^2
mean.temp	Annual mean temperature	CHELSA climate data	1km^2
human.density	Human population density	European Commission	~1km^2, bilinear resampling
friction	Friction map in which each pixel represents the speed of human travel in that area	Weiss *et al.* 2018	1km^2
min.temp	Coldest temperature of the coldest month	CHELSA climate data	1km^2
EvBroadleaf	Evergreen/deciduous needleleaf trees (%)	Tuanmu and Jetz 2014	1km^2
Needleleaf	Needleleaf trees (%)	Tuanmu and Jetz 2014	1km^2
DecBroadleaf	Deciduous broadleaf trees (%)	Tuanmu and Jetz 2014	1km^2
MiscTrees	Mixed/other trees (%)	Tuanmu and Jetz 2014	1km^2
Shrubs	Shrubs (%)	Tuanmu and Jetz 2014	1km^2
Herb	Herbaceous vegetation (%)	Tuanmu and Jetz 2014	1km^2
Crop	Cultivated and managed vegetation (%)	Tuanmu and Jetz 2014	1km^2
Flood	Regularly flooded vegetation (%)	Tuanmu and Jetz 2014	1km^2
Urban	Urban/built-up (%)	Tuanmu and Jetz 2014	1km^2
Snow	Snow/ice (%)	Tuanmu and Jetz 2014	1km^2
Barren	Barren including spare shrub/herbaceous cover (%)	Tuanmu and Jetz 2014	1km^2
Water	Open water (%)	Tuanmu and Jetz 2014	1km^2
Slope	Slope	Amatulli *et al.* 2020	90m^2, resampled by taking mean of pixels in 1km^2
Altitude	Altitude	MERIT DEM	90m^2, resampled by taking mean of pixels in 1km^2
PET	Monthly potential evapotranspiration	CGIAR CSI	1km^2
DailyTempRange	Mean diurnal temperature range	CHELSA climate data	1km^2
max.temp	Maximum temperature of the warmest month	CHELSA climate data	1km^2
AnnualTempRange	Annual temperature range	CHELSA climate data	1km^2
prec.wet	Precipitation of the wettest month	CHELSA climate data	1km^2
prec.dry	Precipitation of the driest month	CHELSA climate data	1km^2
GPP	Gross primary production, a measure of vegetation photosynthesis	Zhang *et al.* 2017	500m^2, bilinear resampling
kernel100	Kernel density map of sampled sites (bandwidth = 100km)	NA	1km^2

Table S3. Description and equations for each performance metric recorded for the model. The equations reference the randomforestSRC package in R and variables are defined as follows: RF = Random Forest model under consideration, GD = genetic distance measure (CSE or linearized F_{ST}), TestingData = predictor and observational data from 1 site and affiliated pairs, TrainingData = predictor and observational data from the other 37 sites.

Metrics for 10-fold cross-validation		
Abbreviation	**Description**	**Equation**
RSQ	Psuedo R-squared (percent variance explained by model, calculated by R package)	1 - mse / Var(y)
$RMSE_{train}$	Root mean square error of model for training dataset	sqrt(mean((RF$predicted.oob - TrainingData$FST_lin)^2))
$RMSE_{test}$	Root mean square error of model for testing dataset	sqrt(mean((predict.rfsrc(RF, TestingData)$predicted - TestingData$GD)^2))
MAE_{train}	Mean absolute error of model for training dataset	mean(abs(predict.rfsrc(RF, TrainingData)$predicted - TrainingData$GD))
MAE_{test}	Mean absolute error of model for testing dataset	mean(abs(RF$predicted.oob - TestingData$GD))
R_{train}	Pearson correlation between predicted and observed genetic distance for training dataset	cor(RF$predicted.oob,TrainingData$GD)
R_{test}	Pearson correlation between predicted and observed genetic distance for testing dataset	cor((predict.rfsrc(RF, TestingData))$predicted, TestingData$GD)
Metrics for full dataset run		
Abbreviation	**Description**	**Equation**
RSQ_{full}	Psuedo R-squared (percent variance explained by model, calculated by R package)	1 - mse / Var(y)
$RMSE_{full}$	Root mean square error of model	sqrt(mean((RF$predicted.oob - FullData$CSE)^2))
R_{full}	Pearson correlation between predicted and observed genetic distance for full dataset	cor(RF$predicted.oob, FullData$GD)
MAE_{full}	Mean absolute error of model for full dataset	mean(abs(RF$predicted.oob - FullData$GD))

Table S4. Leave-one-out cross-validation results for iterative Random Forest model using CSE as genetic distance. **Best iteration** selected by highest R_{test}; **RSQ_{train}** = R-squared (percent variance explained by the model); **$RMSE_{train}$** = root mean squared error of model for training dataset; **$RMSE_{test}$** = root mean squared error of model for validation dataset; **MAE_{train}** = mean absolute error of model for training dataset; **MAE_{test}** =mean absolute error of model for validation dataset; **R_{train}** = correlation between observed and predicted CSE using training dataset; **R_{test}** = correlation between observed and predicted CSE using testing dataset; **Most Important variables** are the four most important variables for optimized Random Forest model. For detailed information about these metrics, see Table S5.

Point	Site	It.	RSQ	$RMSE_{train}$	$RMSE_{test}$	MAE_{train}	MAE_{test}	R_{train}	R_{test}	Most important variables
1	Az	0	0.618	0.039	0.042	0.030	0.035	0.708	0.708	Max temp, Altitude, Human density, Kernel 100
2	TJC2	7	0.661	0.036	0.030	0.029	0.024	0.834	0.834	Max temp, Human density, Slope, Mean temp
3	Nog	10	0.662	0.036	0.033	0.028	0.025	0.860	0.860	Max temp, Slope, Barren, Human density
4	Her	2	0.671	0.036	0.036	0.028	0.027	0.797	0.797	Max temp, Slope, Altitude, Mean temp
5	Dem	6	0.674	0.036	0.028	0.028	0.021	0.828	0.828	Max temp, Slope, Human density, Mean temp
6	LC18	10	0.685	0.035	0.026	0.027	0.019	0.867	0.867	Max temp, Access, Altitude, Mean temp
7	TC	6	0.680	0.035	0.050	0.027	0.040	0.453	0.453	Max temp, Slope, Access, Barren
8	Soc	6	0.663	0.035	0.049	0.027	0.040	0.633	0.633	Max temp, Slope, Access, Human density
9	Alb	5	0.667	0.035	0.047	0.027	0.037	0.529	0.529	Max temp, Slope, Herb, Human density
10	Ala	0	0.611	0.039	0.038	0.031	0.032	0.750	0.750	Max temp, Altitude, Human density, Kernel 100
11	SP	9	0.667	0.036	0.035	0.028	0.026	0.832	0.832	Max temp, Slope, Access, Herb
12	Ros	6	0.703	0.034	0.074	0.026	0.068	0.655	0.655	Max temp, Access, Mean temp, Slope
13	Car	2	0.685	0.035	0.034	0.027	0.029	0.661	0.661	Max temp, Slope, Altitude, Mean temp
14	Lov	7	0.685	0.035	0.026	0.028	0.020	0.750	0.750	Max temp, Mean temp, Altitude, Herb
15	Lub	0	0.601	0.040	0.034	0.031	0.028	0.572	0.572	Max temp, Altitude, Human density, Kernel 100
16	Dall	9	0.672	0.036	0.037	0.028	0.030	0.720	0.720	Max temp, Slope, Access, Mean temp
17	El	1	0.654	0.037	0.032	0.028	0.026	0.748	0.748	Max temp, Altitude, Human density, Kernel 100
18	Cam	2	0.685	0.035	0.045	0.027	0.039	0.870	0.870	Max temp, Slope, Mean temp, Barren
19	Tr	0	0.604	0.039	0.040	0.031	0.034	0.813	0.813	Max temp, Altitude, Kernel 100, Human density
20	Bex	6	0.658	0.037	0.033	0.029	0.027	0.789	0.789	Max temp, Slope, Human density, Barren
21	Houston	2	0.680	0.035	0.078	0.028	0.070	0.454	0.454	Max temp, Mean temp, Altitude, Human density
22	Nuec	8	0.668	0.036	0.038	0.028	0.032	0.704	0.704	Max temp, Slope, Altitude, Access
23	Hid	9	0.665	0.036	0.053	0.028	0.045	0.729	0.729	Max temp, Access, Altitude, Barren

24	Juar	5	0.663	0.036	0.041	0.028	0.033	0.780	0.780	Max temp, Slope, Barren, Access
25	NL	9	0.657	0.035	0.088	0.028	0.077	0.423	0.423	Max temp, Mean temp, Altitude, Herb
26	NO	2	0.666	0.036	0.053	0.028	0.043	0.415	0.415	Max temp, Slope, Altitude, Mean temp
27	18/9	6	0.691	0.035	0.039	0.027	0.033	0.607	0.607	Max temp, Mean temp, Slope, Altitude
28	StA	2	0.678	0.035	0.038	0.027	0.033	0.921	0.921	Max temp, Slope, Mean temp, Altitude
29	Bb	0	0.600	0.039	0.036	0.031	0.032	0.938	0.938	Max temp, Altitude, Kernel 100, Human density
30	DB	5	0.665	0.036	0.018	0.028	0.015	0.964	0.964	Max temp, Slope, Herb, Barren
31	Orl	1	0.656	0.036	0.034	0.027	0.029	0.980	0.980	Max temp, Mean temp, Altitude, Access
32	Mel	3	0.670	0.035	0.021	0.028	0.017	0.942	0.942	Max temp, Slope, Herb, Mean temp
33	FLO	10	0.656	0.036	0.024	0.028	0.020	0.972	0.972	Max temp, Slope, Access, Barren
34	PBC	7	0.656	0.036	0.021	0.028	0.018	0.960	0.960	Max temp, Slope, Access, Mean temp
35	Miami	1	0.651	0.036	0.023	0.028	0.020	0.949	0.949	Max temp, Friction, Human density, Kernel100
36	FM	8	0.662	0.036	0.023	0.028	0.018	0.961	0.961	Max temp, Slope, Barren, Herb
37	Sar	4	0.674	0.035	0.022	0.027	0.018	0.963	0.963	Max temp, Slope, Mean temp, Herb
38	Tam	9	0.642	0.037	0.019	0.029	0.015	0.968	0.968	Max temp, Slope, Access, Mean temp
Mean			0.661	0.036	0.038	0.028	0.031	0.771	0.771	
Stdev			0.025	0.001	0.016	0.001	0.014	0.169	0.169	

Table S5: Mean relative importance of all variables for the leave-one-out cross-validation using CSE as genetic distance.

Variable	Mean relative importance
Maximum temperature	1.000
Slope	0.495
Altitude	0.343
Mean temperature	0.313
Human density	0.269
Accessibility to nearest major city	0.268
Kernel 100	0.268
Herbaceous vegetation	0.259
Barren	0.224
Shrubs	0.181
Precipitation of the wettest month	0.128
Daily temperature range	0.122
Gross primary production	0.116
Friction	0.116
Deciduous broadleaf	0.113
Potential Evapotranspiration	0.113
Annual precipitation	0.111
Annual Temperature Range	0.109
Precipitation of the driest month	0.095
Aridity	0.095
Water	0.091
Needleleaf	0.090
Minimum temperature	0.088
Urban	0.087
Evergreen broadleaf	0.085
Crop	0.082
Regularly flooded vegetation	0.065
Misc. trees	0.065
Snow	0.008

Table S6. Leave-one-out cross-validation results for iterative Random Forest model using linearized F_{ST} as genetic distance. **Best iteration** selected by lowest $RMSE_{test}$; **Site** = the site excluded for the training dataset; **RSQ** = Pseudo R-squared (percent variance explained by the model); **$RMSE_{train}$** = root mean squared error of model for training dataset; **$RMSE_{test}$** = root mean squared error of model for testing dataset; **MAE_{train}** = mean absolute error of model for training dataset; **MAE_{test}** =mean absolute error of model for validation dataset; **R_{train}** = correlation between observed and predicted linearized F_{ST} using training dataset; **R_{test}** = correlation between observed and predicted linearized F_{ST} using testing dataset; **Most Important variables** are the four most important variables for optimized Random Forest model. For detailed information about these metrics, see Table S5.

Point	Site	It.	RSQ	$RMSE_{train}$	$RMSE_{test}$	MAE_{train}	MAE_{test}	R_{train}	R_{test}	Most important variables
1	Az	0	0.489	0.031	0.032	0.009	0.032	0.703	0.461	Slope, Mex temp, Kernel100, Friction
2	TJC2	2	0.500	0.030	0.024	0.023	0.019	0.708	0.708	Access, Max temp, Kernel100, Slope
3	Nog	8	0.550	0.027	0.047	0.021	0.036	0.744	0.600	Max temp, Altitude, Access, DecBroadleaf
4	Her	10	0.556	0.029	0.029	0.022	0.023	0.747	0.630	Max temp, Altitude, Access, DecBroadleaf
5	Dem	0	0.474	0.032	0.032	0.011	0.035	0.697	0.699	Slope, Max temp, Altitude, Access
6	LC18	4	0.580	0.028	0.027	0.021	0.021	0.765	0.836	Access, Slope, Kernel100, Max temp
7	TC	5	0.582	0.028	0.037	0.021	0.029	0.766	0.331	Max temp, Mean temp, Access, Kernel100
8	Soc	10	0.572	0.028	0.036	0.021	0.029	0.758	0.638	Max temp, Access, Mean temp, Kernel100
9	Alb	2	0.530	0.029	0.034	0.022	0.027	0.729	0.409	Kernel100, Max temp, Water, Access
10	Ala	8	0.524	0.030	0.027	0.023	0.020	0.727	0.694	Access, Max temp, Kernel100, Mean temp
11	SP	2	0.535	0.030	0.024	0.023	0.021	0.734	0.853	Slope, Access, Max temp, Friction
12	Ros	4	0.553	0.029	0.028	0.022	0.025	0.745	0.757	Access, Max temp, Mean temp, Kernel100
13	Car	6	0.546	0.029	0.024	0.022	0.020	0.740	0.679	Max temp, Mean temp, Access, Slope
14	Lov	1	0.535	0.030	0.026	0.023	0.019	0.733	0.657	Slope, Max temp, Mean temp, Kernel100
15	Lub	0	0.471	0.032	0.024	0.009	0.033	0.689	0.561	Max temp, Slope, Kernel100, Access
16	Dall	6	0.550	0.029	0.021	0.023	0.017	0.743	0.761	Max temp, Slope, Access, Mean temp
17	El	4	0.530	0.030	0.025	0.023	0.021	0.730	0.741	Max temp, Access, Mean temp, Altitude
18	Cam	10	0.563	0.028	0.048	0.021	0.040	0.751	0.596	Max temp, Access, Mean temp, Altitude
19	Tr	9	0.536	0.030	0.020	0.023	0.016	0.733	0.830	Max temp, Slope, Access, Mean temp

20	Bex	8	0.547	0.029	0.022	0.022	0.016	0.741	0.733	Access, Max temp, Altitude, Kernel100
21	Houston	10	0.611	0.026	0.068	0.020	0.059	0.783	0.572	Access, Max temp, Slope, Flood
22	Nuec	8	0.584	0.028	0.026	0.021	0.021	0.766	0.690	Max temp, Slope, Access, Flood
23	Hid	7	0.527	0.030	0.036	0.023	0.030	0.726	0.716	Access, Max temp, Slope, Altitude
24	Juar	9	0.504	0.031	0.024	0.023	0.020	0.710	0.761	Access, Slope, Max temp, Mean temp
25	NL	9	0.540	0.029	0.039	0.022	0.029	0.737	0.506	Access, Max temp, Kernel100, Mean temp
26	NO	10	0.559	0.029	0.044	0.022	0.038	0.750	0.462	Max temp, Kernel100, Access, Mean temp
27	18/9	3	0.563	0.029	0.030	0.021	0.024	0.752	0.668	Access, Slope, Max temp, Mean temp
28	StA	10	0.537	0.030	0.028	0.023	0.024	0.736	0.825	Access, Kernel100, Max temp, Mean temp
29	Bb	5	0.579	0.027	0.056	0.020	0.050	0.762	0.863	Slope, Access, Max temp, Mean temp
30	DB	8	0.487	0.031	0.019	0.024	0.016	0.698	0.920	Access, Max temp, Slope, Kernel100
31	Orl	5	0.508	0.030	0.022	0.023	0.017	0.714	0.941	Mean temp, Access, Max temp, Slope
32	Mel	4	0.551	0.029	0.021	0.022	0.017	0.747	0.865	Slope, Max temp, Flood, Mean temp
33	FLO	8	0.543	0.029	0.012	0.022	0.011	0.739	0.963	Mean temp, Max temp, Access, Altitue
34	PBC	4	0.545	0.029	0.014	0.022	0.011	0.741	0.934	Slope, Max temp, Mean temp, Kernel100
35	Miami	4	0.531	0.030	0.017	0.023	0.014	0.730	0.918	Mean temp, Access, Max temp, Kernel100
36	FM	1	0.544	0.029	0.018	0.022	0.014	0.740	0.907	Max temp, mean temp, Kernel100, Slope
37	Sar	10	0.530	0.030	0.020	0.023	0.016	0.734	0.840	Access, Max temp, Kernel100, Slope
38	Tam	6	0.538	0.030	0.017	0.023	0.013	0.736	0.914	Mean temp, Max temp, Slope, Kernel100
Mean			0.540	0.029	0.029	0.021	0.024	0.736	0.722	
Stdev			0.031	0.001	0.012	0.004	0.010	0.021	0.160	

Table S7: Mean relative importance of all variables for the leave-one-out cross-validation using linearized F_{ST} as genetic distance.

Variable	Mean relative importance
Maximum temperature	0.824
Accessibility to the nearest major city	0.765
Slope	0.555
Mean temperature	0.547
Kernel 100	0.466
Altitude	0.354
Deciduous broadleaf	0.248
Shrubs	0.239
Needleleaf	0.229
Annual temperature range	0.215
Misc. trees	0.198
Friction	0.195
Water	0.189
Evergreen broadleaf	0.188
Daily temperature range	0.187
Precipitation of the wettest month	0.181
Annual precipitation	0.180
Potential evapotranspiration	0.177
Aridity	0.173
Regularly flooded vegetation	0.171
Precipitation of the driest month	0.152
Cultivated and managed vegetation	0.146
Human density	0.144
Barren	0.139
Minimum temperature	0.124
Herbaceous vegetation	0.122
Gross primary production	0.120
Urban	0.086
Snow	0.004

Table S8: Result for full dataset run of iterative Random Forest model using CSE as genetic distance. **Best iteration** selected by lowest $RMSE_{test}$; **Site** = the site excluded for the training dataset; **RSQ** = Pseudo R-squared (percent variance explained by the model); **$RMSE_{full}$** = root mean squared error of model for training dataset; **MAE_{full}** = mean absolute error of model for training dataset; **R_{full}** = correlation between observed and predicted CSE; **Most Important variables** are the four most important variables for optimized Random Forest model. For detailed information about these metrics, see Table S5.

Iteration	**RSQ**	**$RMSE_{full}$**	**MAE_{full}**	**R_{full}**	**Most important variables**
Straight	0.606	0.0388	0.031	0.786	Max temp, Altitude, Human density, Kernel 100
1	0.674	0.0353	0.027	0.825	Max temp, Altitude, Mean temp, Human density
2	0.669	0.0356	0.028	0.824	Max temp, Slope, Altitude, Kernel 100
3	**0.688**	**0.0345**	**0.027**	**0.832**	**Max temp, Slope, Barren, Human density**
4	0.667	0.0357	0.028	0.820	Max temp, Slope, Barren, Herb
5	0.685	0.0347	0.027	0.830	Slope, Max temp, Barren, Herb
6	0.661	0.0360	0.028	0.817	Max temp, Slope, Herb, Barren
7	0.662	0.0359	0.028	0.817	Max temp, Altitude, Herb, Human density
8	0.669	0.0356	0.028	0.821	Max temp, Slope, Human density, Barren
9	0.658	0.0361	0.028	0.818	Max temp, Slope, Human density, Mean temp
10	0.680	0.0349	0.027	0.828	Max temp, Slope, Barren, Human density

Table S9: Comparison of root mean squared error (RMSE) among full model, leave-one-out cross-validation (LOOCV), and leave-two-out cross-validation (LTOCV).

Model	**$RMSE_{train}$ ± SD**	**$RMSE_{test}$ ± SD**
Full model	0.035 (=$RMSE_{full}$)	
LOOCV	0.036 ± 0.0014	0.038 ± 0.016
LTOCV	0.036 ± 0.0016	0.038 ± 0.015

Dataset S1 (separate file). Input dataset for the iterative Random Forest model. Each row represents a pair of sites; the latitude and longitude for each site is listed as well as the genetic distance (CSE and linearized F_{ST}) between each pair of sites.

Dataset S2 (separate file). Microsatellite calls for all individuals in STRUCTURE format.

Dataset S3 (separate file). Most important variables for each iteration of the full dataset run using CSE as genetic dataset.

SI References

[1] B. Peng, M. Kimmal, simuPOP: a forward-time population genetics simulation environment. *bioinformatics* 21(18), 3686-3687 (2005).

[2] B. Peng, C. Amos, Forward-time simulations of nonrandom mating populations using simuPOP. *bioinformatics* 24(11), 1408-1409 (2008).

[3] F. Rousset, genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* 8, 103-106 (2008).