Temperature and population density determine reservoir regions of seasonal persistence in highland malaria

Amir S. Siraj[1,2], Menno J. Bouma[3], Mauricio Santos-Vega[4], Asnakew K. Yeshiwondim[5], Dale S. Rothman[2], Damtew Yadeta[6], Paul C. Sutton[1,7], Mercedes Pascual[4,8].

[1] Department of Geography & the Environment, University of Denver, 235 Boettcher West, 2050 E. Iliff Ave. Denver, Colorado 80208-0710, USA.
[2] Frederick S. Pardee Center for International Futures, Josef Korbel School of International Studies, University of Denver, 2201 South Gaylord Street, Denver, Colorado 80208-0500, USA.
[3] London School of Hygiene and Tropical Medicine, University of London, London WC1 E7HT, UK.
[4] Department of Ecology and Evolution, University of Chicago, 1101 E 57th Street, Chicago, Illinois 60637, USA.
[5] PATH/ Malaria Control and Elimination Partnership in Africa. Africa Avenue, Getu Commercial Center, P.O Box 493, code 1110, Addis Ababa, Ethiopia.
[6] Oromia Regional Health Bureau, P.O. Box 24341, Addis Ababa, Ethiopia.
[7] School of Natural and Built Environments, University of South Australia, P Building, Mawson Lakes Campus, Mawson Lakes SA 5095, Australia
[8] Howard Hughes Medical Institute, Chevy Chase, Maryland 20815-6789, USA.

Sept 16, 2015

Electronic Supplementary Material

1. Data and Methods

Epidemiological data

A monthly time series of *Plasmodium falciparum* cases confirmed through microscopy examination of blood slides from clinical (febrile) individuals is used in this study. The data is from 159 subunits (kebeles) surrounding the Debre Zeit sector, collected at the malaria examination and treatment center in Bishoftu town for the period from September 1993 to February 2007. We have excluded malaria data after August 2005 because of the introduction of a new treatment (ACTs) for *P. falciparum* in September 2005 in the aftermath of the severe epidemic years (2002-2004) in Ethiopia. The attached movie shows the spatial distribution of the number of malaria cases (normalized by population) in each kebele for every month in this time period. In order to look at the dynamics at the seasonal level, we aggregated the monthly data for each kebele into four month seasonal blocks for January-March (JFMA), May-August (MJJA), September-December (SOND), representing respectively the low, intermediate and high transmission seasons (for a total of 11 data points per kebele for each of JFMA, MJJA and SOND seasons).

Demographic and cartographic data

Each of the 159 kebele further encompasses up to 4 smaller administrative sub-units for which population data were obtained from the Central Statistical Agency of Ethiopia for 1994 and 2007 [1-2]. We interpolated these population data temporally based on growth rates between the two censuses conducted in 1994 and 2007, separately considering changes in urban and rural populations at the district level. Spatial coordinates for these sub-units were obtained from the Oromia regional Bureau of Health. These coordinates, along with the population data, were used to weight all spatially explicit variables and obtain population weighted estimates aggregated at the kebele level.

Two estimates of population density were considered. After spatially overlaying the administrative sub-units with their population sizes, we drew two circles of 5 and 10 km radii respectively, around each of these. The sum of the population of the sub-units that fall within each of these circles was divided by the area of the circle to obtain two estimates of population density for each sub-unit The value was calculated separately for each sub-unit, for each 4 month block, since population growth rates may vary depending on which district the sub-unit belongs to. These values were then averaged at the kebele level to obtain two estimates of population density for each kebele.

We obtained the digital elevation model (DEM) at a 30-meters resolution from the Global Earth portal managed by USGS [3]. By using ArcGIS software tools, we generated gridded slopes from the high resolution DEM. For better data manageability, we resampled the high resolution DEM, averaged at a 5-arc-minutes resolution. Finally, we overlaid the locations of our administrative sub-units on the DEM and the slope gridded surfaces to obtain estimates for altitude and slope respectively at the sub-unit level.

Climate Data

Daily readings of minimum and maximum temperature for Ethiopian stations were obtained from the National Meteorological Agency (NMA). Four meteorological stations in Ethiopia situated in close proximity to the study areas (namely Addis Ababa (Bole), Addis Ababa (Obs), Adama and Debre Zeit) were selected based on their proximity to Bisoftu/ Debre Zeit town and their high correlation with the Debre Zeit's station readings. Missing data were filled by estimating from the linear association between altitudes and temperatures, the values for the remaining station. We then developed regional minimum and maximum temperature lapse rates for each month of the year. This was done by pooling all four-station readings for a single month of the year (eg. January) and regressing the temperature readings on the corresponding station altitude values to estimate a slope (the change in temperature for a given change in altitude).

These slopes were used to estimate the minimum and maximum monthly temperature at the altitude of each administrative sub-unit, by assuming the average of the four-station readings corresponds to the average altitude of the four stations (=2071 meters above sea level). Estimates at each administrative sub-unit were weighted by their respective population and averaged at the kebele level.

Our approach of using altitudinal difference as a proxy for temperature difference over space is appropriate for this study because (a) the kebeles have close proximity of stations - all 154 kebeles are within 50 km from a station (5 kebeles have distance between 50 and 57km) (b) all kebeles being within a single rainfall regime -one of a total of 12 in Ethiopia [4] (c) All kebeles in the study are on the same side of a continuous elevation gradient (with the exception of 7 in the far north), constituting part of a single micro climate.

We further assessed temperature readings and locations of the four stations (a) the four sites are appropriately situated on two opposite sides (East and West) and in the middle of study area, providing good spatial spread of sample data for interpolation (see Fig. S2 B) (b) single year DJF mean temperature for the four sites are highly associated with altitude (R-squared ranging from 0.88 to 0.99 with median 0.95 for all 11 seasons). (c) year to year DJF mean temperature have high associations across sites (R-squared ranging from 0.84 to 0.96 for the all combinations of two stations).

Moreover, in seasons with stable low moisture and thus pressure levels, relative humidity decreases strongly as temperature increases. Since, we are focusing on the dry period of the year, during which the moisture level in the entire region is uniformly at its lowest level, we expect relative humidity to be driven by temperature and not vise-versa. Therefore, we believe the effect of humidity on temperature in the months of DJF will be minimal, and thus differences in altitude can be reasonable approximations to differences in temperature.

Daily readings of rainfall from 13 stations in close proximity to Bishoftu town were obtained from the National Meteorological Agency (NMA). These stations are Addis Ababa (Bole), Addis Ababa (Obs), Aleltu, Chefe Donsa, Debre Zeit, Dertu Liben, Ejere, Guranda Meta, Hombole (had only 9 months of data), Koka Dam, Mojo, Nazeret and Sebeta. Missing data were filled by randomly selecting one reading from the same date of the year in the four nearby years: two years prior and two years after. If missing data were for Feb 29th of a leap year, we randomly selected from the Feb 28th data for the four years. Daily readings from these stations were then spatially interpolated by using ordinary Kriging [5]. We cross-validated model estimates at each station by varying the type of variogram model and its parameters (sill, nugget and range). We selected the best global parameters from a range of reasonable values by comparing the root mean square errors for all stations [6]. The interpolated grids were constructed to have a spatial resolution of 0.5 degrees and to cover the entire study region. Administrative sub-unit coordinates were then overlaid to obtain (drill down) daily estimates, which were then aggregated at the kebele level weighted by the sub-unit population.

Finally, all daily estimates of rainfall and temperature were aggregated at one, two and three month blocks to examine associations with cases in the low transmission season.

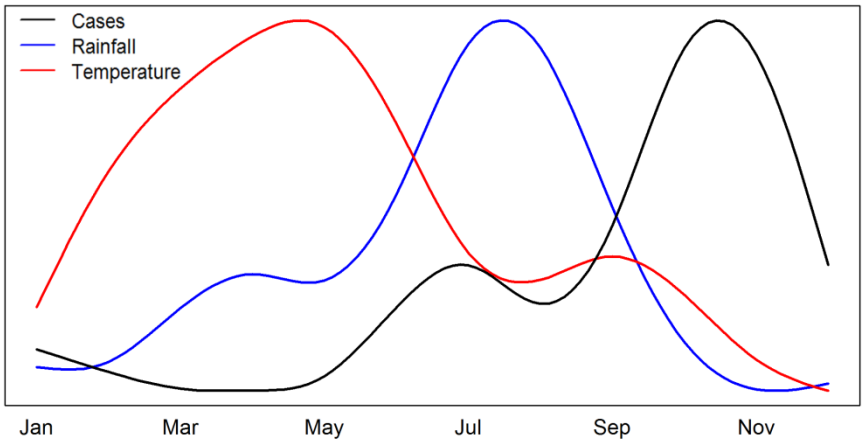

Figure S1: Annual cycle for cases (black line), rainfall (blue line) and temperature (red line) for our study area.

Monthly Sea Surface Temperature (SST) anomalies for the Niño 3.4 region were obtained from NOAA Optimal Interpolation SST Version 2 database [7]. Monthly average Normalized Difference Vegetation Index (NDVI) at a resolution of 0.1 degrees were also obtained from the IRI analysis of USGS data for the period 1993 to 2004 [8]. For 2005, we obtained NDVI from the MODIS/Terra Vegetation Indices for the year 2005 [9] at a temporal resolution of 16 days and spatial resolution of 500m, which were aggregated to obtain monthly averages at a resolution of 0.1 degree which matches those preceding 2005..

We used shape files of perennial rivers obtained from Food and Agriculture Organization [10], and lakes obtained from the Environmental Systems Research Institute [11], and computed the distance between each administrative sub-unit and its closest perennial water body. This distance was considered to have a decay effect (the effect on malaria transmission decays exponentially as distance increases); thus we used the inverse-square distance.

In addition, to examine the ability of local soils to retain rain water (or water holding capacity), we used the GAEZ soil database [12] for the dominant soil type at a resolution of 5 arc minutes. We also obtained ISRIC-WISE soil water content data (in mm) at 30 arc minutes resolution [13], which includes relative size of different soil types and their specific water capacities. By matching the higher resolution dominant soil types layer to the water capacity layer sorted by area size, we were able to obtain water capacity at higher resolution of 5 arc minutes. We then overlaid the administrative sub-units coordinates on the water holding capacity layer to estimate values at each subunit.

Least-cost distance estimates

A least-cost measure is based on the notion that some landscapes are more difficult and costly to traverse than others [14]. is based on the concept of minimizing the accumulated cost of a traveler in moving from point A to point B within a landscape. If the cost of traveling a certain distance is equal for all directions, then the least-cost distance is the Euclidean distance. The concept has been well developed in transportation, environmental economics and archeology, where economic and social forces have a recognized influence on the spatial distribution of events [15-16]. In relation to infectious disease dynamics, the concept applies to capturing differences in connectivity between regions resulting from social and economic activities. We capture these differences by incorporating differences in travel infrastructure types including trails, the latter being of particular importance during the dry season.

In this study, we assume cost of travel is contingent on the kind of surface, namely paved roads, gravel roads or walking trails. Studies have compared costs of travel for paved and dirt roads [15], as well as for paved roads and hiking trails [17]. We recognize that in addition to taking a longer time, walking is more laborious as compared to traveling on vehicles that utilize roads. While the cost of these transportation means is context specific (depending on availability, the quality of roads, the opportunity cost of walking etc.), we use these different categories as reference points for constructing our own assumptions. We specifically assume that for a given distance, travel on gravel roads is 1.5 times costlier than that on paved roads, while walking is two times costlier. Although this estimate seems conservative, it is more likely to be closer to reality than an approach lacking any concept of cost and implying equal cost per distance for all means of mobility.

Neighborhood Structures

We plotted all possible routes of connectivity between each administrative sub-units (n=342) using geographic information system (GIS) methods. We identified the least-cost routes for each pair of administrative sub-units using the ArcGIS´s Network Analyst tool. Then neighborhood structures were implemented by considering 1) adjacent kebeles where neighborhood consists of kebeles with common borders (2) kebeles with sub units at most 5km (paved road distance equivalent) apart (3) kebeles with sub units at most 10km apart (paved road equivalent).

2. Supplementary results:

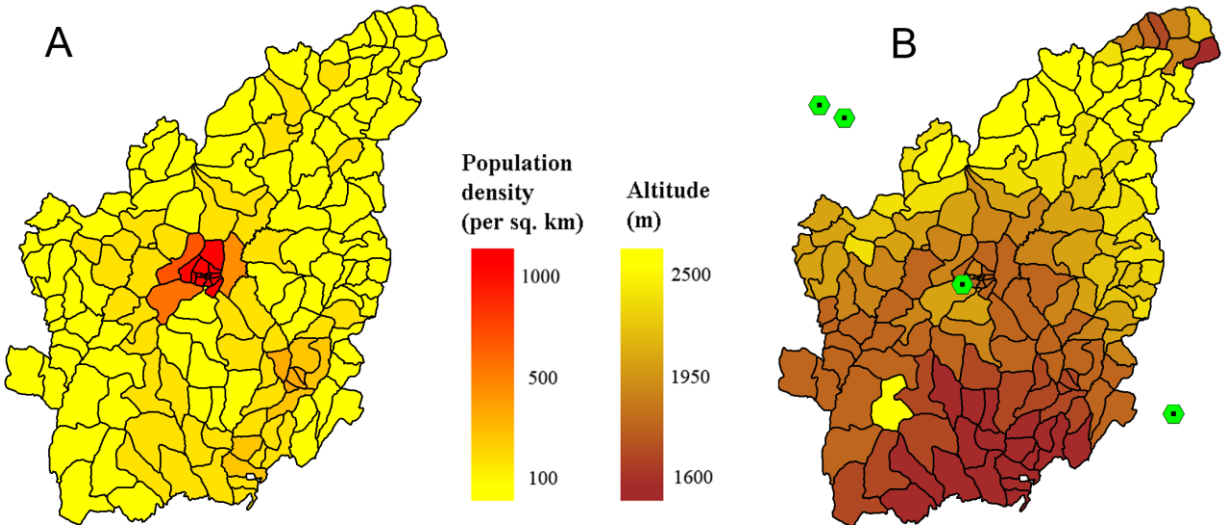

Figure S2: Figure A shows the population density obtained by adding all populations within a 5km radius around each administrative sub-unit (up to 4 per kebele), and by dividing the value by the area of the circle (see text for details). The high-density area at the center corresponds to the Bishoftu/ Debre zeit town (A). Figure B shows the elevation map, with elevation weighted by the population size of the administrative sub-units within each kebele (see text for details). The green points indicate the location of the four meteorological stations used to interpolate mean temperature based on elevation of sub-units in each kebele. The four meteorological stations are Addis Ababa Bole, Addis Ababa Obs, Adama and Debre Zeit.

Table S1: Distribution of kebeles by observed and predicted quantiles of cases

		Observed				
		No case	**Very low**	**Low**	**High**	**Very High**
predicted	**No case**	883	113	27	13	2
	Very low	163	85	47	35	16
	Low	23	32	26	28	7
	High	6	15	16	43	26
	Very high	5	10	8	32	88

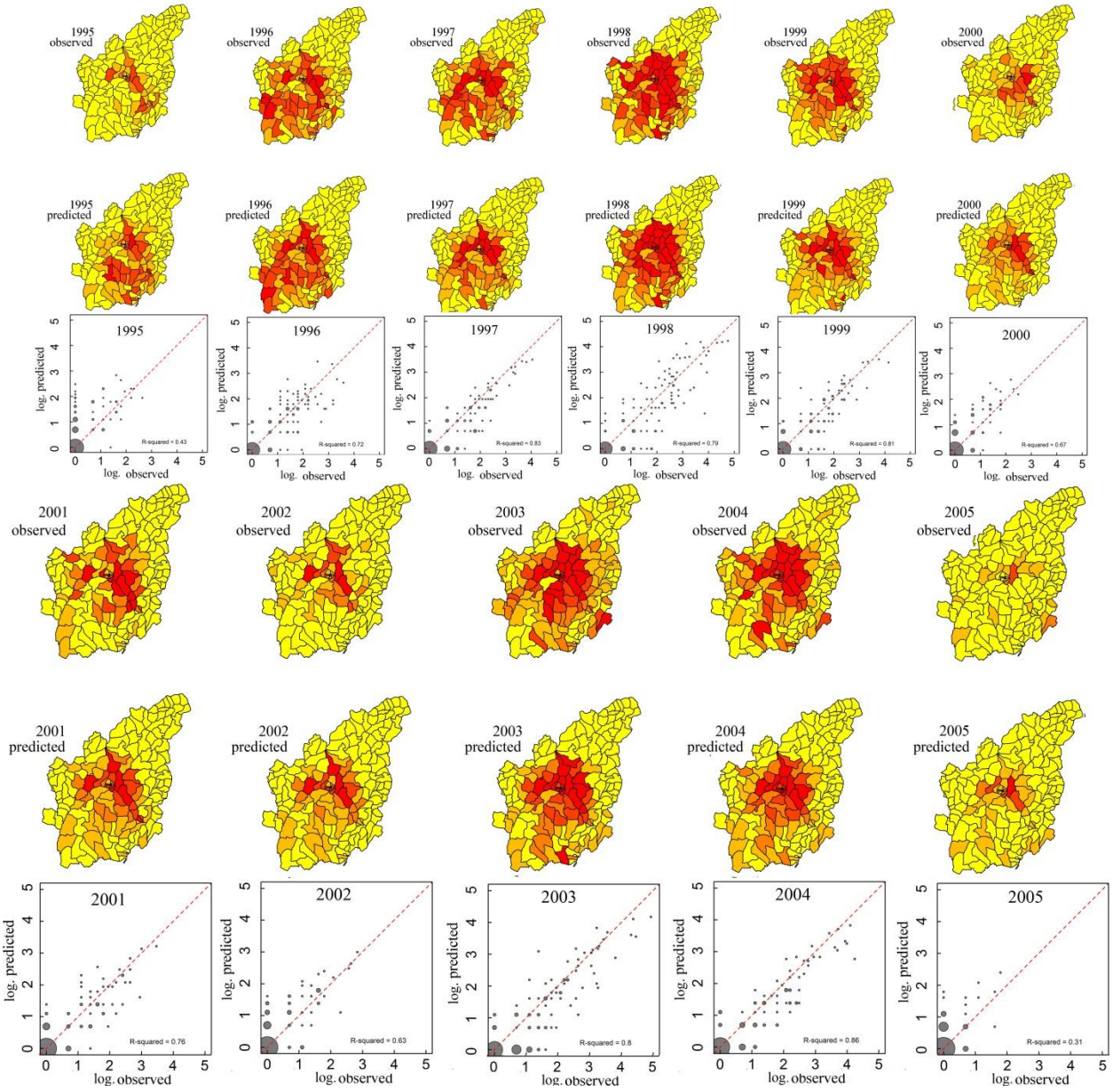

Figure S3: Model predictions are compared to observations for all 11 years based on the best GLMM model (that includes structured and unstructured random effects). The top two rows of panels show maps for the predicted and observed quantiles respectively. The quantiles were generated by considering zero cases in one class, and subdividing all nonzero JFMA cases into four equally-sized intervals, with the resulting categories representing respectively no cases, very low, low, high and very high cases, and the corresponding colors ranging from yellow to red. The lower row shows scatter plots (C and F) for predicted against observed cases (in logarithmic scale), which include the identity line for comparison (in red). The size of each circle in the scatter plots is scaled by the square root of the number of predicted-observed pairs at that point, with the highest number obtained for the (0, 0) pair

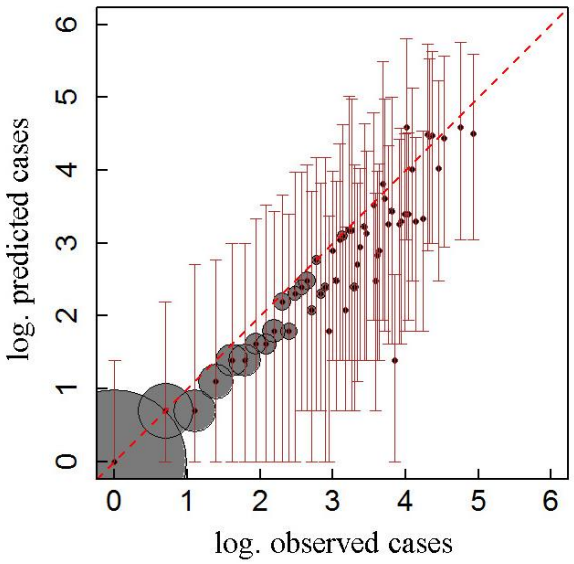

Figure S4: Comparison of predictions (y-axis) and observations (x-axis) for the GLMM model that includes structured and unstructured random effects. Cases for individual kebeles and seasons are shown in log scale. As for Figure 3, predictions are obtained for 10,000 simulations of the model with parameters sampled from the posterior distribution of estimated parameters. Here, a distribution is obtained for all the predictions that correspond to a given observed value. The black dots represent the medians of these distributions and the red vertical intervals, the 95 Credible Intervals. Each dot is surrounded by a circle whose size is scaled by the square root of the number of predictions at that observed value). For example, the highest number (n=1080) is obtained for (0, 0), and numbers decrease with incidence. Median predictions for the most part fall along the identity line (in red) with a slight tendency to under-predict at the high end of the cases. The exception is one outlier (x about 4) whose CI does not straddle the diagonal, which represents a single kebele and a single season (1998). This kebele (Ejersa) is found in the south-east corner of our region, at low altitudes for the region but in a typically dry area. Its proximity to a lake and its location along a major road might explain our under-prediction for this warm year.

Table S2: Coefficients of the best GLMM model for the high transmission season (SOND) cases

Covariate		Median	95% CI	R-hat
Total JJA rainfall	$\beta 1$	-0.1188	[-0.1654, -0.0729]	1.001
Mean JJAS temperature	$\beta 2$	0.6750	[0.5853, 0.7665]	1.001
Population density	$\beta 3$	-0.1681	[-0.4000, 0.0738]	1.009
Indoor Residual Spraying	$\beta 4$	-0.2810	[-0.4727, -0.0806]	1.001
Lagged malaria relative risk	γ	0.2934	[1.3400, 1.5183]	1.001
Spatial unstructured hyper-parameter	σ^2_ϕ	0.0115	[0.0003, 0.1419]	1.163
Spatial structured hyper-parameter	σ^2_v	4.1288	[2.9638, 5.7176]	1.003
Over dispersion parameter	θ^{-1}	1.9180	[1.7000, 2.1600]	1.001

Credible Intervals (CI) obtained from the 2.5% and 97.5% quantiles of the distribution.

Model results based on a subset of the data

The following results are based on a model trained based on a subset of the data consisting of the first six year only.

Table S3: Coefficients of the best model for the cases in the low transmission season (JFMA) fitted to six years of data (1995-2000). The results show rainfall is no longer significant at the 0.05 level, but all other variables and in particular population density remain significant..

Covariate		**Median**	**95% CI***	**R-hat**
Total DJF rainfall	$\beta 1$	-0.0773	[-0.056, 0.213]	1.009
Mean DJF temperature	$\beta 2$	0.4073	[0.260, 0.556]	1.009
Population density	$\beta 3$	0.1604	[0.088, 0.233]	1.002
Lagged malaria relative risk	γ	1.2870	[1.192, 1.396]	1.001
Spatial unstructured hyper-parameter	σ^2_ϕ	0.0097	[0.001, 0.099]	1.020
Spatial structured hyper-parameter	σ^2_v	0.4077	[0.181, 0.792]	1.003
Over dispersion parameter	θ^{-1}	3.1895	[2.403, 4.344]	1.002

** this model was trained by using a subset of the dataset by taking the first six years of data.
* Credible Intervals (CI) obtained from the 2.5% and 97.5% quantiles of each parameter's distribution.

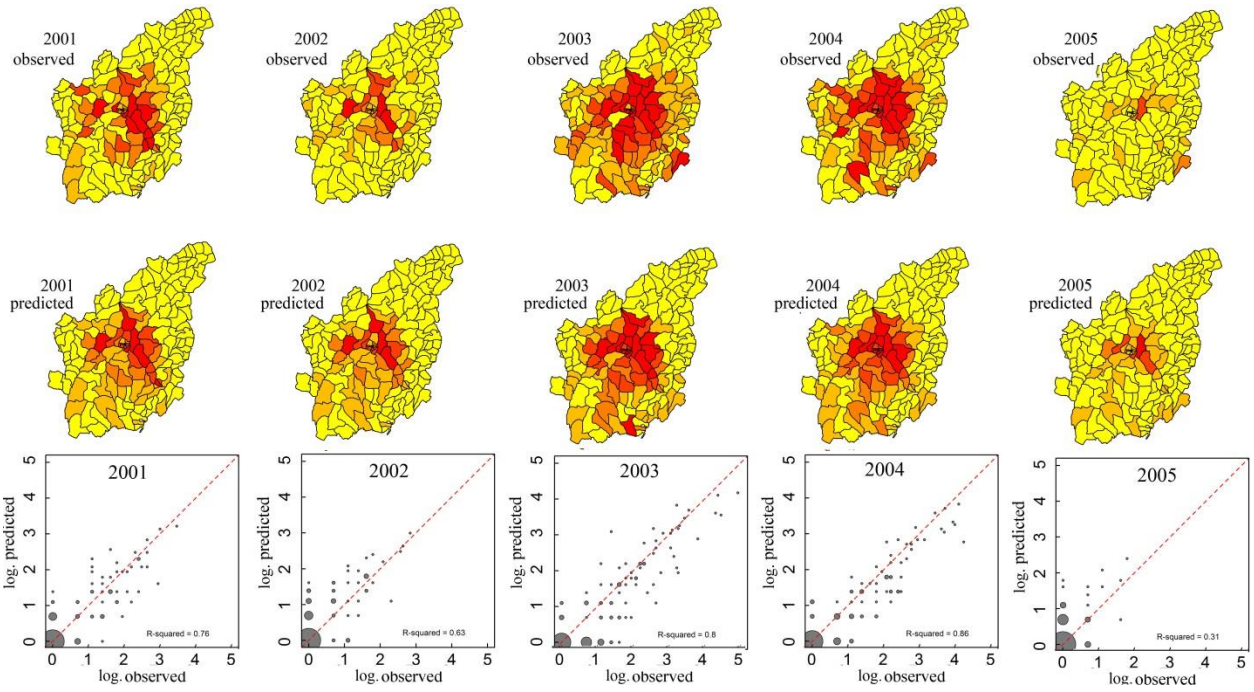

Figure S5: Model predictions are compared to observations for all five years based on the best GLMM model (that includes structured and unstructured random effects). The model was fitted using the first 6 years of data (1995-2000) only, and predictions presented here are for the remaining, "out-of-fit" years, post-2000 The top two rows of panels show maps for the predicted and observed quantiles respectively. The quantiles were generated by considering zero cases in one class, and subdividing all nonzero JFMA cases into four equally-sized intervals, with the resulting categories representing respectively no cases, very low, low, high and very high cases, and the corresponding colors ranging from yellow to red. The lower row shows scatter plots (C and F) for predicted against observed cases (in logarithmic scale), which include the identity line for comparison (in red). The size of each circle in the scatter plots is scaled by the square root of the number of predicted-observed pairs at that point, with the highest number obtained for the (0, 0) pair.

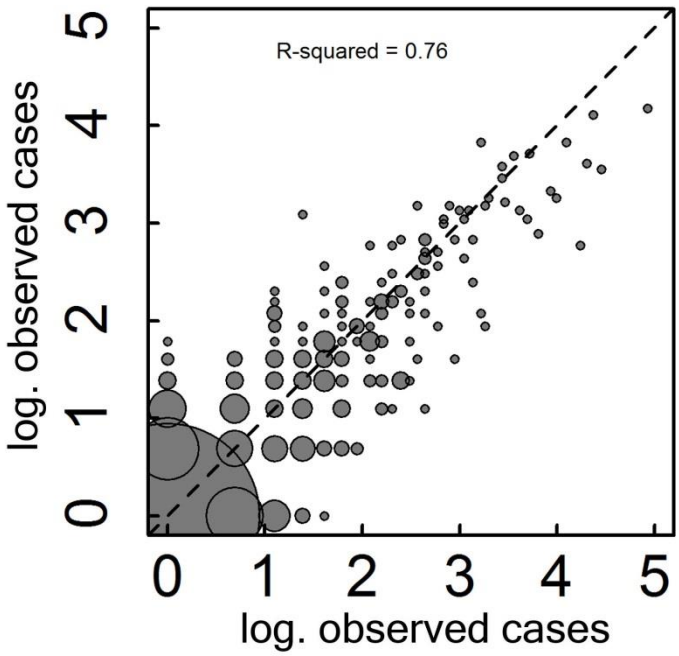

Figure S6: Observed Vs median of fitted JFMA cases (in log scale) with the best best GLMM model (that includes structured and unstructured random effects) rained with data from the first 6 years (1995-2000) only. Each dot represents a single kebele's JFMA cases over the 5 years (2001-2005) not included in the model training. The size of each circle in the scatter plots is scaled by the square root of the frequency at that point, with the highest frequency (n=439) obtained at the (0, 0) point. Note that the years for which these analysis are done were not part of the dataset used in model training.

References

1. CSA, Central Statistical Authority 1996 The 1994 population and housing census of Ethiopia. Results for Oromia Region. Volume I. Addis Ababa, Ethiopia: *Central Statistics Authority*.
2. CSA, Central Statistical Authority 2008 The 2007 population and housing census of Ethiopia. Statistical Report for Oromia Region. Addis Ababa, Ethiopia: *Central Statistics Authority*.
3. USGS's Earth Explorer Online System, accessed on 14 March 2011. http://earthexplorer.usgs.gov/ ASTER GDEM is a product of METI and NASA.
4. Korecha D, Sorteberg A. 2013. Construction of Homogeneous Rainfall Regimes for Ethiopia. In *Characterizing the Predictability of Seasonal Climate in Ethiopia*, PhD Thesis: University of Bergen, 125:176.
5. Krige, DG 1951 A statistical approach to some basic mine valuation problems on the Witwatersrand, *J. Chem. Metall. Min. Soc. S. Afr.* **52**, 119–139.
6. Hofstra N, Haylock M, New M, Jones P, Frei C 2008 Comparison of six methods for the interpolation of daily, European climate data. *Journal of Geophysical Research*, **113**, D21110
7. IRI (International Research Institute for Climate and Society) 2014a Data Library. http://iridl.ldeo.columbia.edu/SOURCES/.Indices/.nino/.EXTENDED/.NINO34/ Accessed on 15 March 2014.
8. IRI (International Research Institute for Climate and Society) 2014b Data Library. http://iridl.ldeo.columbia.edu/expert/SOURCES/.IRI/.Analyses/.USGS/.ADDS/.NDVI/.NDVIg/.deg0p1/.monthly/.c8204/.avgNDVImonavg/ , Accessed on 15 March 2014.
9. USGS (United States Geological Survey) 2014a Modis Vegetation Indices 16-day L3 Global 500m. obtained from site https://lpdaac.usgs.gov/data_access maintained by the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls , South Dakota, Accessed on 12 Sept 2014.
10. FAO (Food and Agriculture Organization of the United Nations). 2014. FAO GEONETWORK. Perennial Water Courses (Rivers) of the World (Vmap0) (GeoLayer). (Latest update: 18 Feb 2014) Accessed on 13 Apr 2015. URI: http://data.fao.org/ref/c0c0dfa0-88fd-11da-a88f-000d939bc5d8.html?version=1.0
11. ESRI (Environmental Systems Research Institute). 2013. *World Lakes. Data & Maps for ArcGIS*. ESRI. Redlands, USA.
12. FAO/IIASA 2011 Global Agro-ecological Zones (GAEZv3.0). FAO, Rome, Italy and IIASA, Laxenburg, Austria.
13. Batjes NH 2005 ISRIC-WISE Global dataset of derived soil properties on a 0.5 by 0.5 degree grid (version 3.0). Report 2005/ 05, ISRIC – World Soil Information, Wageningen (with dataset).

14. Storfer A, Murphy MA, Evans JS, Goldberg CS, Robinson S , Spear SF, Dezzani R, Delmelle E, Vierling L, Waits LP. 2007 Putting the 'landscape' in landscape genetics. *Heredity*, **98**, 128–142.
15. Stone, S.W. 1998 Using a geographic information system for applied policy analysis: the case of logging in the Eastern Amazon. *Ecological Economics*. **27**, 43–61.
16. White, DA; Surface-Evans, SL (2012 *Least Cost Analysis of Social Landscapes: Archaeological Case Studies*. University of Utah Press.
17. Pingel, TJ. 2010 Modeling Slope as a Contributor to Route Selection in Mountainous Areas. *Cartography and Geographic Information Science*, **37**, 137-148.