SUPPLEMENTAL INFORMATION 1: MODEL-BASED GEOSTATISTICAL PROCEDURES

Below are the details of the model-based geo-statistical (MBG) framework used to develop the malaria risk maps for Malawi. The MBG procedure was adopted from Gething and others,[1] where further model details, especially on age-standardization of parasite rate data, can also be found. The generic model code is available on Malaria Atlas Project *Plasmodium falciparum* Cartographic Code Link[2] and has been adapted for Malawi. Model fitting was achieved by using Markov chain Monte Carlo (MCMC).[3,4]

SI 1.1 MBG presentation.

Each of the N_i individuals in sample i was assumed *P. falciparum* positive with probability $\tilde{k}_i P'(x_i, t_i)$, so the number positive N_i^+ was distributed binomially:

$$N_i^+ | N_i, P'(x_i, t_i) \overset{\text{ind}}{\sim} \text{Bin}\left(N_i, \tilde{k}_i P'(x_i, t_i)\right) \quad \text{S1.1}$$

The coefficient $P'(x_i, t_i)$ was modeled as a Gaussian process. The factor $\tilde{k}_i$ converted $P'(x_i, t_i)$ to the probability that individuals within the age range reported for study i were *P. falciparum* positive, and that the infection was detected, thereby accounting for the influence of age on the probability of detection.[5] The age-standardization factor $\tilde{k}_i$ in each population was assumed drawn independently from a distribution $D_{\tilde{k}}$ whose parameters were the lower $A_{L,i}$ and upper $A_{U,i}$ ages reported in study i:

$$\tilde{k}_i | A_{U,i}, A_{L,i} \overset{\text{ind}}{\sim} D_{\tilde{k}}(A_{U,i}, A_{L,i}) \quad \text{S1.2}$$

The form of $D_{\tilde{k}}$ is described by Gething and others.[1]

Pf$\text{PR}_{2\text{–}10}$ is the *P. falciparum* parasite rate for individuals between ages 2 (2.00) and 10 years (9.99). Its value at an arbitrary location x and time t is the product of $P'(x, t)$ and another age-standardization factor, $k_{2\text{–}10}$, distributed as $D_k(2, 10)$:

$$\begin{aligned} &\text{PR}_{2-10}(x, t) = P'(x, t) k_{2-10}(x, t) \\ &k_{2-10}(x, t) \overset{\text{ind}}{\sim} D_k(2, 10) \end{aligned} \quad \text{S1.3}$$

The factor $k_{2\text{–}10}$ converted $P'(x, t)$ to the probability that individuals between ages 2 and 10 years at location x are *P. falciparum* positive. The age-standardization factor $\tilde{k}$ of a survey is the product of the age-standardization factor k associated with the same place, time and age range and the sensitivity of the survey.

The coefficient $P'(x, t)$ at arbitrary location x and time t was modeled as the inverse-logit function applied to a random field f evaluated at (x, t), plus an unstructured (random) component $\epsilon(x, t)$.

$$P'(x, t) = \text{logit}^{-1}(f(x, t) + \epsilon(x, t)) \quad \text{S1.4}$$

The components $\epsilon(x, t)$ were assumed independent and identically distributed for each location x and time t and a standard diffuse but proper prior with expectation 0.25 was assigned to their variance V.

$$\epsilon(x, t) | V \overset{\text{iid}}{\sim} \text{N}(0, V) \quad \text{S1.5}$$

$$\frac{1}{\gamma} \sim \textit{Gamma}(3, 12) \quad \text{S1.6}$$

The random field f was modeled as a Gaussian process characterized by its mean and covariance functions:

$$f(x, t) | \beta, \tau, \phi_x, \phi_t, \lambda, \psi, \rho, \upsilon \sim \text{GP}(\beta, C) \quad \text{S1.7}$$

The mean function was defined as $\mu = \beta \mathbf{X}$,where $\mathbf{X} = 1$, $X_1(x), \ldots, X_n(x)$ was a vector consisting of a constant and $N = 2$ environmental covariates indexed by spatial location x ,and $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_n$ was a corresponding vector of regression coefficients. The covariance of the field was modeled by using a version of the spatiotemporal covariance function recently recommended by Stein[6] (equation 2.12):

$$\begin{aligned} &C(x_i, t_i; x_j, t_j) = \tau^2 \gamma(0) \frac{(\Delta x)^{\gamma(\Delta t)} K_{\gamma(\Delta t)}(\Delta x)}{2^{\gamma(\Delta t)-1} \Gamma(\gamma(\Delta t)+1)}, \\ &\gamma(\Delta t) = \frac{1}{2\rho + 2(1-\rho)[(1-\upsilon) e^{-|\Delta t|/\phi_t} + \upsilon \cos(2\pi \Delta t)]}, \\ &\Delta t = |t_i - t_j| \end{aligned} \quad \text{S1.8}$$

K_γ is the modified Bessel function of the second kind of order γ, and Γ is the gamma function [7,8].

Spatial distance between a pair of points x_i and x_j was computed as great-circle distance $D_{GC}(x_i, x_j)$ multiplied by a factor that depends on the angle of inclination $\theta(x_i, x_j)$ of the vector pointing from x_i to x_j. θ was computed as if latitude and longitude were Euclidean coordinates (on a cylindrical projection) to allow for anisotropy:

$$\Delta x = 2\sqrt{\gamma(\Delta t)} \frac{D_{GC}(x_i, x_j) \sqrt{1 - \psi^2 \cos^2(\theta(x_i, x_j) - \lambda)}}{\phi_x} \quad \text{S1.9}$$

When $\Delta x = 0$ (that is, for points at the same location but different times), the covariance function reduces to

$$\rho + (1-\rho)\left[(1-\upsilon) e^{-|\Delta t|/\phi_t} + \upsilon \cos(2\pi \Delta t)\right] \quad \text{S1.10}$$

As temporal separation increases, the covariance approaches a limiting sinusoid $\tau^2[\rho + (1 - \rho)\upsilon \cos(2\pi\Delta t)]$ rather than zero. When $\Delta t = 0$, on the other hand (for points at different locations but the same time), it reduces to a standard exponential form with range parameter $\phi_x\sqrt{2}$. Unlike standard sum-product models, this covariance function does not have problematic ridges along its axes.[6]

SI 1.2 Prior specification.

The square root of the partial sill τ and the spatial range parameter ϕ_x were assigned skew-normal priors:

$$\log \tau | \mu_\tau, V_\tau, \alpha_\tau \sim \text{Skew-Normal}(\mu_\tau, V_\tau, \alpha_\tau) \quad \text{S1.11}$$

$$\log \phi_x | \mu_\phi, V_\phi, \alpha_\phi \sim \text{Skew-Normal}(\mu_\phi, V_\phi, \alpha_\phi) \quad \text{S1.12}$$

and their specification is described further below.

The standard one-over-x prior for the temporal scale parameter ϕ_t resulted in collapse to 0, a common artifact when data do not contain strong information. A relatively vague but proper prior, which has an expectation of ten years, was used instead.

$$\phi_t \sim \text{Exponential}(0, .1) \quad \text{S1.13}$$

A uniform prior was assigned to the direction of anisotropy parameter λ and to the square of the eccentricity parameter ψ, which controls the amount of anisotropy,

$$\lambda \sim \text{Uniform}(0, \pi) \quad \text{S1.14}$$

$$\psi^2 \sim \text{Uniform}(0, 1) \quad \text{S1.15}$$

a uniform prior was assigned to the temporal parameters governing the amplitude of the sinusoidal component ρ and the limiting autocorrelation in the temporal direction υ:

$$\rho \sim \text{Uniform}(0, 1), \upsilon \sim \text{Uniform}(0, 1) \quad \text{S1.16}$$

and a standard prior was assigned to the components of the mean:

$$p(\beta) \propto 1 \quad \text{S1.17}$$

Although standard priors such as the improper flat prior[3] were assigned to most of the basic model parameters, subjective skew-normal priors[7] were specified for the range and partial sill parameters τ and ϕ_x.

SI 1.3 Model implementation.

SI 1.3.1 MCMC algorithms.

The main geostatistical model and the age-standardization sub-model were fitted by using the MCMC algorithm.[3,4] The algorithm was implemented in the Python[8] and FORTRAN programming languages by using the open-source Bayesian statistics package PyMC[9,10] and the numerical packages SciPy and NumPy.[11]

The evaluation of f at the sampling locations and times was updated by using Gibbs steps.[3] The evaluation of the uncorrelated process ϵ was updated one point at a time by using random-walk Metropolis steps. [3] The model parameters β, τ, ϕ_x, ϕ_t, λ, ψ, V, and ρ were updated jointly by using the method of Haario and others.[12]

Within the MCMC loop, the age-standardization factors $\tilde{k}_i$ were not imputed explicitly. We were not interested in their particular values, and marginalizing out nuisance parameters ahead of time usually improves the mixing of MCMC algorithms. Before the MCMC loop began, the marginal likelihood:

$$\int Bin\left(N_i^+; N_i, k_i P'(x_i, t_i)\right) D_{\tilde{k}}(\tilde{k}_i; A_{U,i}, A_{L,i}) d\tilde{k}_i \quad \text{S1.18}$$

was approximated by using standard Monte Carlo integration for several values of $P'(x_i, t_i)$. That is, values for the model parameters α_i, b_i, c_i, and s_i and the age distribution s_i were drawn from their posterior predictive distributions, then expression (S1.3) was evaluated to obtain k_i, then the binomial probability was evaluated for several values of $P'(x_i, t_i)$. The probabilities resulting from many such draws were averaged. Inside the MCMC loop, the marginal likelihood function for arbitrary values of $P'(x_i, t_i)$ was evaluated by interpolation.

SI 1.3.2 Age correction model.

The age distribution parameters S_i, S_0 and v are independent of the relative *Pf*PR parameters P_i', α_i, c_i, b_i, s_i, μ_A, σ, and R given the data, so these two groups of parameters were inferred by using separate MCMC algorithms.

In the MCMC for the age distribution parameters, the survey populations' age distributions s_i were updated by using Gibbs steps.[3] The concentration parameter V was updated by using random-walk Metropolis steps.[3] The typical age distribution S_0 was represented as a normalized sequence of gamma random variables,[13] and these variables were updated one at a time by using random-walk Metropolis steps.[3]

In the MCMC for the relative *Pf*PR parameters, the distributional parameters μ_A, σ and R were updated jointly by using the method of Haario and others.[12] The parameters P_i', α_i, c_i, b_i, and s_i were updated jointly for each population i by using the same method.

SI 1.3.3 Spatiotemporal prediction.

The output of the MCMC stage consisted of $\{\theta_{(l)}; l = 1, \ldots, m\}$ samples from the posterior of the parameter set $\theta = \{\beta, \tau, \phi_x, \phi_t, \lambda, \psi, \rho, k, V\}$ and a corresponding $\{f(x_i, t_i)_{(l)}; l = 1, \ldots, m\}$ samples from the posterior of the space-time random field at each of the n data locations $\{(x_i, t_i); i = 1, \ldots, n\}$. For every $l'th$ sample, the conditional distribution of the annual mean of the space-time random field was predicted at each prediction location x_j on the nodes of a regular 1 × 1 km grid within the spatial limits of stable *P. falciparum* transmission.[14] The distribution of the annual mean $f(x_j)_{(l)}$ for prediction location x_j was modeled as the joint multivariate normal distribution of the 12 predicted monthly values e.g., ($l = 2010_{tan}, \ldots, 2010_{Oss}$) for that year specified by a 12-element mean vector $\hat{\boldsymbol{y}}(x_j)_{(l)}$ and 12 × 12 variance-covariance matrix $\hat{\boldsymbol{\sigma}}^2(x_j)_{(l)}$:

$$f(x_j)_{(l)} \sim MVN\left(\hat{\boldsymbol{y}}(x_j)_{(l)}, \hat{\boldsymbol{\sigma}}^2(x_j)_{(l)}\right) \quad \text{S1.19}$$

The mean vector $\hat{\boldsymbol{y}}(x_j)_{(l)}$ was computed by using

$$\hat{\boldsymbol{y}}(x_j)_{(l)} = \boldsymbol{\mu}_{P_{(l)}} + \boldsymbol{C}_{DP_{(l)}}^T \cdot \boldsymbol{C}_{DD_{(l)}}^1 \cdot \left(\boldsymbol{p}(x, t) - \boldsymbol{\mu}_{D_{(l)}}\right) \quad \text{S1.20}$$

where $\boldsymbol{\mu}_P$ and $\boldsymbol{\mu}_D$ were the predicted mean of the random field at each of the 12 prediction times ($l = 2010_{tan}, \ldots, 2010_{Oss}$) at spatial location x_j and at each of the n data locations respectively, $\boldsymbol{C}_{DP}$ and $\boldsymbol{C}_{DD}$ were the data-to-prediction and data-to-data covariance matrices respectively, and $\boldsymbol{p}(x, t)$ was the vector of n data values. The 12 × 12 variance-covariance matrix $\hat{\boldsymbol{\sigma}}^2(x_j)_{(l)}$ was computed by using

$$\hat{\boldsymbol{\sigma}}^2(x_j)_{(l)} = \boldsymbol{C}_{PP_{(l)}} - \boldsymbol{C}_{DP_{(l)}}^T \cdot \boldsymbol{C}_{DD_{(l)}}^{-1} \cdot \boldsymbol{C}_{DP_{(l)}} \quad \text{S.1.21}$$

The value of the $l'th$ sample of V, the variance of the unstructured component $\epsilon(x, t)$, was then added to the diagonal of the matrix $\hat{\boldsymbol{\sigma}}^2(x_j)_{(l)}$ and 1,000 draws were made randomly from the distribution specified in equation S1.19 These draws represented samples from the posterior distribution of $f(x_j)$ and were subject to an inverse logit transform and then multiplied by the $l'th$ sample of the age-standardization parameter $k_{2-10(l)}$ to form the $l'th$ sample from the posterior distribution of the predicted mean annual 2000, 2005, and 2010 *Pf*PR_{2-10} endemicity surfaces at location x_j:

$$P'_{2-10}(x_j)_{(l)} = \text{logit}^{-1}\left(f(x_j)_{(l)} + \epsilon_{(l)}\right) k_{2-10(l)} \quad \text{S.1.22}$$

This procedure was repeated for every $l'th$ sample to form the set $\left\{P'_{2-10}(x_j)_{(l)}; l=1, \ldots, m\right\}$ of m samples for each prediction location. The point estimate of *Pf*PR_{2-10} endemicity at each location was defined as the mean of this set, and the

probability of membership to each class was computed as the proportion of these samples falling within each class definition.

SUPPLEMENTAL REFERENCES

1. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IR, Johnston GL, Tatem AJ, Hay SI, 2011. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J 10:* 378, doi:10.1186/1475-2875-10-378.
2. Malaria Atlas Project, *P. falciparum Cartographic Code*. Available at: https://github.com/malaria-atlas-project/mbgw-clean. Accessed December 2011.
3. Gilks WR, Spiegelhalter DJ, 1999. *Markov Chain Monte Carlo in Practice. Interdisciplinary Statistics*. Boca Raton, FL: Chapman and Hall/CRC Press LLC.
4. Gelman A, Carlin JB, Stern HS, 1993. *Bayesian Data Analysis. Texts in Statistical Science*. Boca Raton, FL: SA: Chapman and Hall/CRC Press LLC.
5. Smith DL, Guerra CA, Snow RW, Hay SI, 2007. Standardizing estimates of the *Plasmodium falciparum* parasite rate. *Malar J 6:* 131.
6. Stein ML, 2005. Space-time covariance functions. *J Am Stat Assoc 100:* 310–321.
7. Azzalini A, 1985. A class of distributions which includes the normal ones. *Scand J Stat 12:* 171–178.
8. van Rossum G, 2008. *Python Programming Language - Official Website*. Available at: http://www.python.org. Accessed July 14, 2013.
9. Fonnesbeck C, Huard D, Patil AP, *PyMC 2.0 User's Guide: Installation and Tutorial*, 2008. Available at: http://www.trichech.us/pymc.
10. Patil A, Huard D, Fonnesbeck CJ, 2010. PyMC: Bayesian stochastic modelling in Python. *J Stat Softw 35:* e1000301.
11. Oliphant TE, 2007. Python for scientific computing. *Comput Sci Eng 9:* 10–20.
12. Haario H, Saksman E, Tamminen J, 2001. An adaptive metropolis algorithm. *Bernoulli 7:* 223–242.
13. Hogg RV, Craig AT, 2005. *Introduction to Mathematical Statistics*. Upper Saddle River, NJ: Prentice Hall Inc.
14. Guerra CA, Gikandi PW, Tatem AJ, Noor AM, Smith DL, Hay SI, Snow RW, 2008. The limits and intensity of *Plasmodium falciparum* transmission: implications for malaria control and elimination worldwide. *PLoS Med 5:* e38.

SUPPLEMENTAL INFORMATION 2: COVARIATE COEFFICIENT ESTIMATES

TABLE SI 2.1

Univariate analysis of considered model covariates, Malawi*

Covariate	Estimate	SE	Z	*P*
TSI	1.7218	0.4942	3.484	< 0.001
Urban	–0.74397	0.18345	–4.055	< 0.001
Precipitation	0.0001277	0.0003238	0.394	0.69343
EVI	0.07315	1.53206	0.048	0.9619

*TSI = temperature suitability index; EVI = enhanced vegetation index.

TABLE SI 2.2

Covariates selected by total sets analysis for inclusion in prediction model, Malawi

Covariate	Estimate	SE	t	*P*
Intercept	0.2545488	0.02657700	9.577785	< 0.001
TSI	0.2751015	0.06463496	4.256234	< 0.001
Urban1	–0.1227187	0.02157419	–5.688218	< 0.001

*TSI = temperature suitability index.

SUPPLEMENTAL INFORMATION 3: POPULATION-ADJUSTED PREVALENCE BY DISTRICT

TABLE SI 3.1

Population-adjusted prevalence (%) by district, 2000 and 2005, Malawi*

	2000	2005
District	PA*Pf*PR$_{2-10}$	PA*Pf*PR$_{2-10}$
Northern region		
Chitipa	36.8	36.7
Karonga	41.6	41.4
Mzimba	33.8	33.7
Nkhata Bay	39.9	39.7
Rumphi	35.7	35.6
Central region		
Dedza	36.0	35.9
Dowa	36.8	36.7
Kasungu	37.7	37.6
Lilongwe	31.3	31.1
Mchinji	39.2	39.1
Nkhotakota	41.0	40.9
Ntcheu	39.0	38.9
Ntchisi	37.5	37.4
Salima	40.5	40.4
Southern region		
Balaka	40.3	40.1
Blantyre	24.9	24.8
Chikwawa	42.4	42.3
Chiradzulu	37.2	37.0
Machinga	41.4	41.3
Mangochi	40.9	40.7
Mulanje	39.4	39.3
Mwanza	40.0	39.9
Neno	41.7	41.5
Nsanje	41.4	41.2
Phalombe	40.1	40.0
Thyolo	39.0	38.8
Zomba	36.6	36.5
Total	36.4	36.3

*PA*Pf*PR$_{2-10}$ = population-adjusted *Plasmodium falciparum* rate.