medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

INTEGRATING GEOSTATISTICAL MAPS AND TRANSMISSION MODELS USING ADAPTIVE MULTIPLE IMPORTANCE SAMPLING

By Renata Retkute[*,†], Panayiota Touloupou[‡], Maria-Gloria Basanez[§], T. Deirdre Hollingsworth[¶] and Simon E.F. Spencer[‡]

University of Cambridge[†], University of Warwick[‡], Imperial College London[§] and University of Oxford[¶]

The Adaptive Multiple Importance Sampling algorithm (AMIS) is an iterative technique which recycles samples from all previous iterations in order to improve the efficiency of the proposal distribution. We have formulated a new statistical framework based on AMIS to sample parameters of transmission models based on high-resolution geospatial maps of disease prevalence, incidence, or relative risk. We tested the performance of our algorithm on four case studies: ascariasis in Ethiopia, onchocerciasis in Togo, HIV in Botswana, and malaria in the Democratic Republic of the Congo.

1. Introduction. Geostatistical modelling has been applied to map a range of infectious diseases at high spatial resolution and multinational scale; examples comprise: malaria (1), soil-transmitted helminthiasis (2), leishmaniases (3), onchocerciasis (4), dengue (5), human African trypanosomiasis (6), HIV (7), and diphtheria-pertussis-tetanus vaccine coverage (8). These maps are made by averaging over many spatially continuous surfaces, which are constructed using geo-positioned survey data, ecological covariates and spatial correlations (9; 10; 11).

Transmission dynamic models have successfully been utilized to evaluate the population dynamics of infectious diseases and assess the impact of interventions (12). It is a common practice to estimate transmission model parameters based on geographically located data, but geostatistical maps provide a unique opportunity to parameterize transmission models at national or even continental scales. However, it is challenging to explore the whole parameter space efficiently, especially when many possible parameter combinations can produce similar values of model output. Generating a large number of parameter sets and running simulations based on these parameters would require substantial computational resources.

[*]To whom correspondence should be addressed.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

In this work we develop a statistical framework based on the Adaptive Multiple Importance Sampling algorithm (AMIS; (13)) for effective integration of geostatistical maps and transmission models. We investigate the performance of AMIS on four case studies: the soil-transmitted nematode *Ascaris lumbricoides* (causative of ascariasis, a soil-transmitted helminth) in Ethiopia; the filarial parasite *Onchocerca volvulus* (causative of onchocerciasis or river blindness) in Togo; HIV infection in Botswana, and infection by the protozoan *Plasmodium* parasite (causative of malaria) in the Democratic Republic of the Congo (DRC). The results show that the proposed framework can successfully be applied for integrating geostatistical maps and transmission models. The resulting combined output constitutes a geographical projection illustrating how the map will evolve through time as well as how the algorithm can be extended to sample parameters in the presence of multiple sampling times or post-control data.

2. Background and methods. Integrating high-resolution geostatistical maps and disease transmission models requires exploring the whole parameter space efficiently, especially when many possible parameter combinations can produce similar values of model output. We propose a statistical framework based on the Adaptive Multiple Importance Sampling algorithm (13) to target important and under-explored areas of the parameter space based on high-resolution nation-wide maps of infection prevalences.

2.1. *Adaptive Multiple Importance Sampling.* Suppose that we need to sample from a complex target distribution π. Importance sampling is based on using weighted samples from a proposal distribution Φ, i.e. $\theta_j \sim \Phi$. The corresponding importance weights are $w_j = \frac{\pi(\theta_j)}{\phi(\theta_j)}$, where $\pi(\theta)$ and $\phi(\theta)$ are the target and the proposal density functions (14).

Veach and Guibas (15) proposed the *deterministic multiple mixture* to pool together importance samples from different proposal distributions. In this case, the importance samples $y_j^t \sim \Phi_t$, $(1 \leq t \leq T,\ 1 \leq j \leq N_t)$ and the corresponding importance weights are calculated based on the mixture of weights (13) given by:

$$w_j^t = \frac{\pi(\theta_j^t)}{\frac{1}{N_1+\cdots+N_t}\sum_{l=1}^{t} N_l \phi_l(\theta_j^t)}. \tag{2.1}$$

This idea was extended by Cornuet and co-authors (13) who proposed using the deterministic multiple mixture formula to construct importance proposals sequentially and adaptively. Firstly, the importance weights for current

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

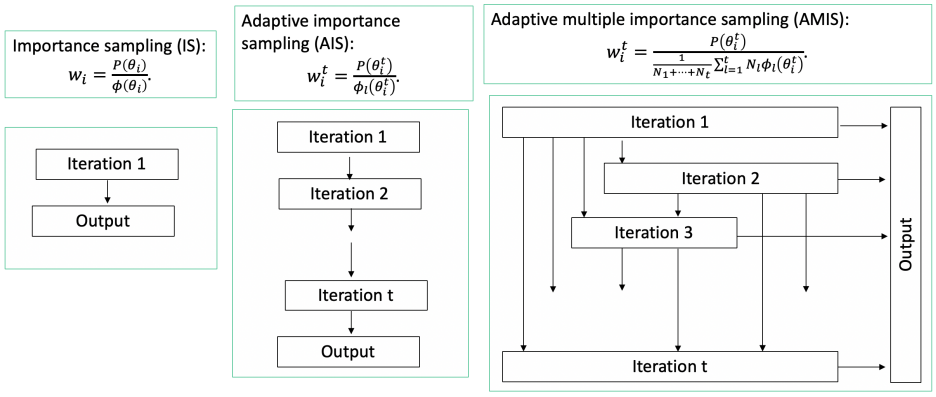


FIG 1. *Diagrammatic illustration of differences between importance sampling (left column), adaptive importance sampling (middle column) and adaptive multiple importance sampling (right column).*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

iteration, t, are calculated, while the importance weights for previous iterations $1 \leq u \leq t-1$ are re-calculated, based on all proposals up to the iteration t. Second, samples from all iterations are used to construct the next proposal distribution.

Figure 1 shows differences between importance sampling (IS), adaptive importance sampling (AIS) and adaptive multiple importance sampling (AMIS). In the case of IS, we can reconstruct the target from a proposal distribution by re-weighting sampled parameters according to importance weights, but it may require a lot of samples to have a reasonably good reconstruction. A more efficient scheme is AIS, where samples from previous iterations are used to construct the proposal distribution. As discussed above, AMIS recycles all samples from previous iterations to both construct the proposal for the current iteration, as well as to reconstruct the target distribution.

Adaptive multiple importance sampling has been utilised in a variety of research fields, including population genetics (16), environment illumination computations (17), and signal communications (18).

The novelty of our study is: (i) applying the AMIS approach to real-world geospatial data; (ii) applying AMIS to multiple targets via the same proposal and (iii) working with a 'moving' target which changes with each iteration of AMIS. As there is a lack of studies on integrating geostatistical maps and transmission dynamics epidemiological models, we believe that the proposed framework will be a valuable addition to the literature from both theoretical and practical perspective.

2.2. *Iterative sampling based on a geostatistical map.* We assume that a map has I pixels (or grid cells), and each pixel has M draws of characteristics in which we are interested, e.g. infection prevalences, annual incidence or number of cases. From here onwards, we will assume that we are dealing with the prevalence of infection. Suppose we have an observed prevalence matrix $Q = (q_{i,m})_{I \times M}$, where each row represents a pixel, and each column represents a sampled surface from a geostatistical map.

Projections of infection prevalence can be quantified using a mathematical model (19; 20). Suppose the we have a mathematical model which we define as $F(\theta)$, and this model translates the parameter space onto the prevalence space $[0, 1]$ with individual parameter vector θ_j corresponding to prevalence p_j, i.e. $F(\theta_j) = p_j$.

At iteration $t = 1$ parameters are sampled from a prior distribution, then the transmission model is used to calculate the prevalence corresponding to the sampled parameters. Parameter vector j is weighted for each pixel i so that the weighted distribution of simulated prevalences resembles the

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Algorithm AMIS for integrating geostatistical maps and transmission models.

1: Set $t = 1$ and let $A = \{1, \ldots, I\}$ be the set of active pixels.

2: Sample N_t parameter vectors from proposal density $\phi_t(\theta)$ and label them θ_j, for $j = \sum_{u=1}^{t-1} N_u + 1, \ldots, \sum_{u=1}^{t} N_u$. If $t = 1$, ϕ_t is a prior distribution $\phi_1(\theta) = \pi(\theta)$. If $t > 1$, ϕ_t is the density of a mixture of Student's t-distributions fitted to $\left\{(\theta_j, \bar{w}_j^{t-1}) : j = 1 \ldots (N_1 + \cdots + N_{t-1})\right\}$.

3: Simulate model: $p_j = F(\theta_j)$.

4: Calculate weights for parameter vectors $j = 1 : (N_1 + \cdots + N_t)$, for each pixel with index $i \in \{1, \ldots, I\}$:

$$w_{ij}^t = \frac{f_i(p_j)}{g(p_j)} \frac{\pi(\theta_j)}{\frac{1}{N_1+\cdots+N_t} \sum_{u=1}^{t} N_u \phi_u(\theta_j)},$$

where $f_i(p_j)$ and $g(p_j)$ are proportional to the probability density of having prevalence p_j under the geostatistical model and the simulation model, respectively, and so f/g is proportional to the Radon-Nikodym derivative. Both probability densities are unavailable and must be estimated using an empirical alternative given by the height of the histogram bar of width δ centred at p_j (21):

$$f_i(p_j) = \sum_{m=1}^{M} \mathcal{I}\left(p_j - \delta/2 \leq q_{i,m} \leq p_j + \delta/2\right),$$

$$g(p_j) = \sum_{k=1}^{N_1+\ldots+N_t} \frac{\mathcal{I}\left(p_j - \delta/2 \leq p_k \leq p_j + \delta/2\right) \pi(\theta_k)}{\frac{1}{N_1+\cdots+N_t} \sum_{u=1}^{t} N_u \phi_u(\theta_k)}.$$

Here $\mathcal{I}$ is the indicator function.

5: Normalise weights for each pixel $i \in A$:

$$\widehat{w}_{ij}^t = \frac{w_{ij}^t}{\sum_{k=1}^{N_1+\ldots+N_t} w_{ik}^t}.$$

6: Calculate effective sample size (ESS) for each pixel $i \in A$:

$$ESS_i^t = \left(\sum_{j=1}^{N_1+\ldots+N_t} \left(\widehat{w}_{ij}^t\right)^2 \right)^{-1}.$$

7: Remove from A any pixels that meet the required effective sample size ($ESS_i^t \geq ESS^R$).

8: Calculate the average weight of parameter vectors based on individual weights of the remaining pixels:

$$\bar{w}_j^t = \frac{1}{|A|} \sum_{m \in A} \widehat{w}_{m,j}^t.$$

9: Set $t = t + 1$.

10: Repeat from step 2 until A is empty or the maximum number of iterations has been reached.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

distribution of observed prevalences at that pixel. We used Kish's effective sample size (ESS) (22) as a measure of the quality of representation of the pixel's prevalence distribution by the simulations.

At iterations 2+ we remove pixels which have an ESS above the required threshold ESS^R and then construct a new proposal composed of a mixture of p-dimensional Student's t-distributions based on the average weights of the simulations, averaged only over pixels that are yet to reach ESS^R. The intuition is that later proposals will be more suitable for pixels that were not well served by earlier proposals. The t-distribution has heavier tails than the Gaussian distribution giving a more robust importance proposal and providing capacity to capture a wider range of targets (23). We used the R package `mclust` to fit the proposal (24). Iterations continue until all pixels have $ESS_i \geq ESS^R$ or the maximum number of iterations reached. Pseudo code is shown in Algorithm AMIS.

2.3. *Incorporating data from multiple time points.* Our proposed framework can be naturally extended to the case where geospatial maps are available at multiple time points $k = 1 \dots K$. In this case, weights for pixel $i = 1 \dots I$ pixel, and parameter vector $j = 1 : (N_1 + ... + N_t)$ will be equal to:

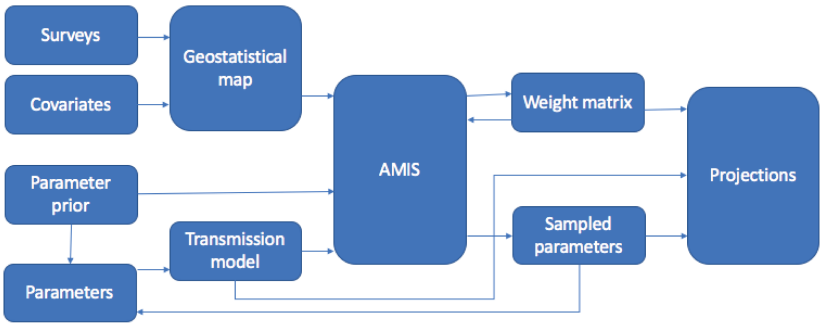
$$w_{i,j}^t = \prod_{k=1}^{K} \frac{f_{i,k}(p_{j,k})}{g_k(p_{j,k})} \frac{\pi(\theta_j)}{\frac{1}{N_1+..+N_t} \sum_{l=1}^{t} N_l \phi_l(\theta_j)}.$$

Again, these weights will favour parameter vectors producing simulated projections which are closer to observed prevalences at multiple time points. This approach can be applied to multiple baseline prevalence maps as well as to the incorporation of post-control maps.

3. Applications. We applied AMIS to four case studies: ascariasis in Ethiopia, onchocerciasis in Togo, HIV in Botswana, and malaria in the DRC. The framework of AMIS approach and description/objectives of each case study are shown in Figure 2.To set up a cse study, three components are required in order to run AMIS framework: geostatistical map, parameter prior distribution and transmission model. We used survey data (25) to produce a series of prevalence maps of *Ascaris lumbricoides* infection with different spatial resolutions and number of observed prevalences at each pixel. For onchocerciasis, HIV, and malaria, we used publicly available maps (4; 7; 1). We have sampled parameters using the AMIS framework and simulated projections for interventions: mass drug administration (MDA) with ivermectin for onchocerciasis in Togo, and the use of insecticide treated nets (ITNs) for malaria in the DRC. For the latter, we focused only on ITNs because the

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

(a)



(b)

Case study	Geostatistical map	Model	Objectives
Ascariasis in Ethiopia	This study	Mathematical relationship	Influence of number of pixels Influence of number of maps Influence of weigh calculation Incorporating dependence between parameters
Onchocerciasis in Togo	O'Hanlon et.al 2016	EPIONCHO (deterministic)	Projections under different control options
HIV infection in Botswana	Dwyer-Lindgren et al. 2019	Deterministic	Incorporating maps at multiple time points Increasing dimension of parameter space
Malaria in the Democratic Republic of the Congo	Bhatt et al. 2015	OpenMalaria (stochastic)	Validation of projections

FIG 2. *(a) AMIS framework. (b) Description of case studies.*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

coverage of artemisinin combination therapies (ACTs) was very low in the DRC (1; 26).

For all four case studies we set $\delta = 0.05$, $N_t = 1000$, $T = 25$, and $ESS_R = 2000$. Therefore the maximum number of sampled parameter vectors was 25,000.

3.1. *Ascariasis in Ethiopia.* *Ascaris lumbricoides* is an intestinal nematode, also known as roundworm (27). It is estimated that 819 million people worldwide are infected by *A.lumbricoides* (25).

3.1.1. *Model and data.* In this study we assumed that the mean number of parasites is a parameter, and we use a simple mathematical relationship between the prevalence and the mean intensity of infection, based on fitting a negative binomial distribution to observed data on worm burdens (19),

$$P = 1 - \left(1 + \frac{W}{k}\right)^{-k}, \tag{3.1}$$

where W is the mean number of worms per host, and k describes the degree of clumping of parasites within a population of hosts (meaning that the distribution of worms per host is overdispersed compared to a Poisson distribution, with k being an inverse measure of overdispersion: the smaller its value, the greater the aggregation of parasites among hosts (28)).

Data on *Ascaris* surveys for Ethiopia were downloaded from the Global Atlas of Helminth Infection (London Applied & Spatial Epidemiology Research Group, LASER) (25). The data consisted of longitude, latitude, number of school-aged children (SAC; 5-14 years old) examined, number of positives (for adult worm presence (25)) and prevalence prior to wide-spread deworming programes (25). Entries with missing geo-location or prevalence values were excluded from further analysis, leaving 290 survey values. We used the INLA-R package (29) and an SPDE approach (30) to produce national level prevalence maps. We incorporated spatial correlation (non-stationary locally isotropic SPDE/GMRF model) and included elevation as an environmental covariate (31) (downloaded from the **raster** package (32)).

3.1.2. *Implementation of AMIS.* We set a uniform prior for log of the mean number of worms $\log(W) \sim U[\log(0.01), \log(60)]$ and a uniform prior for the degree of clumping $k \sim U[0, 3]$. Although typical values for mean worm burden are 10-20 (33), much higher numbers have been observed in field conditions (19; 34).

We produced maps of *Ascaris* prevalence in Ethiopia at resolutions 5km×5km ($I = 37,695$ pixels) and 10km×10km ($I = 11,369$ pixels) and sampled 100,

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

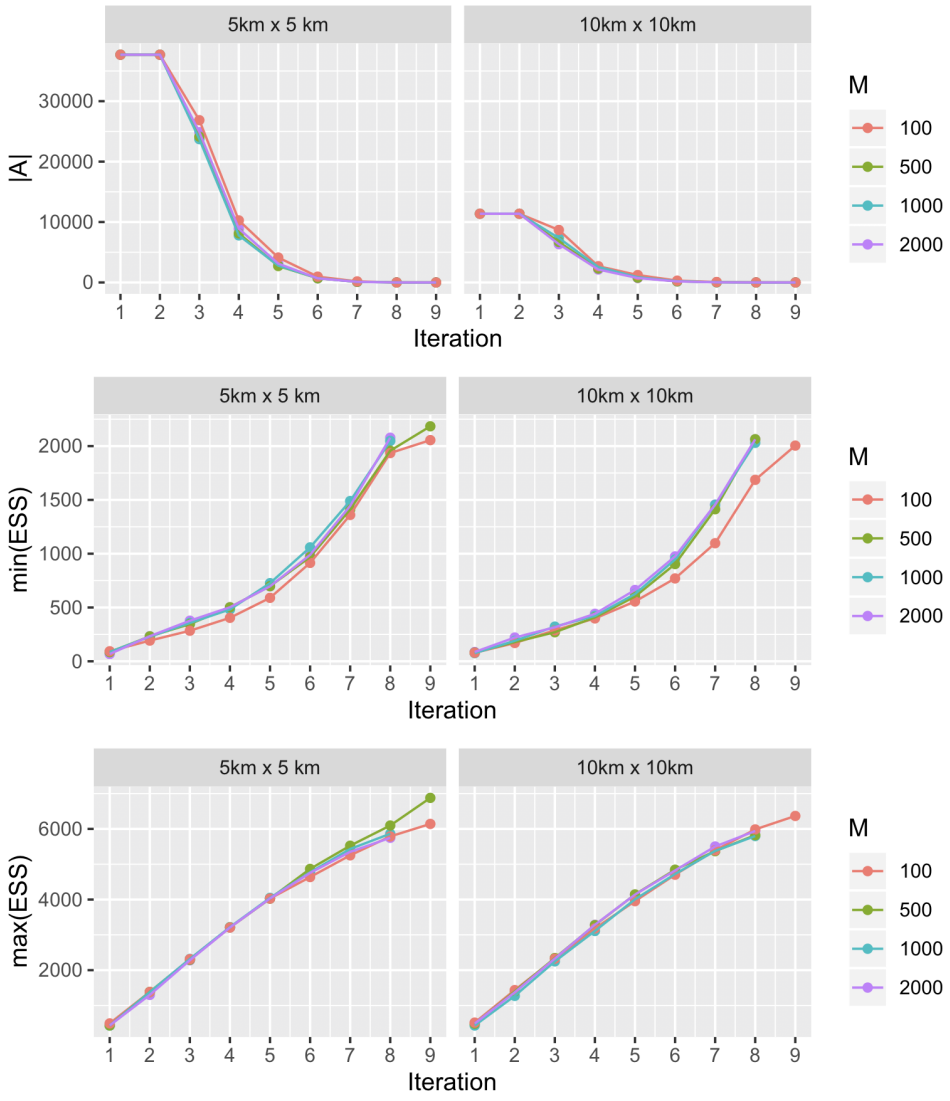


FIG 3. *Change in a number of active pixels, minimum of ESS and maximum of ESS with each AMIS iteration plotted as a function of spatial resolution and number of maps for each pixel in the Ascaris lumbricoides case study.*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

500, 1000 and 2000 individual maps from the posterior distribution for both resolutions. Map showing communities sampled in Ethiopia (25) and triangulated mesh used to build a geostatistical model of *Ascaris* prevalence in Ethiopia (29) are shown in Figure S1.

TABLE 1

Number of iterations to achieve $min(ESS) \geq ESS^R$, with $ESS^R = 2,000$. Here P stands for sampling from the prior only, A_1 stands for AMIS with $|A| = M$ for all iterations, and A_2 is based on algorithm 1. Each iteration samples 1000 parameter sets.

	$M = 100$			$M = 500$			$M = 1,000$			$M = 2,000$		
I	P	A_1	A_2	P	A_1	A_2	P	A_1	A_2	P	A_1	A_2
11,369	23	24	9	21	21	8	22	21	8	21	21	8
37,695	24	18	9	22	18	9	22	18	8	22	17	8

3.1.3. *Results.* We compared our proposed framework against an algorithm which samples parameter vectors at each iteration from the prior only, and against a modified AMIS where we set that $|A| = M$ for all iterations, i.e. algorithm without removing pixels which have achieved $ESS \geq ESS^R$.

The numbers of samples required to meet the ESS target for all pixels are given in Table 1. It can be seen that the adaptive version of AMIS (i.e. with set of active pixels updated every iteration) outperformed other two methods and required around a third of sampled parameters. Eight iterations were required to achieve $\min(ESS) \geq ESS^R$ for resolution 10km×10km and $M \geq 500$, corresponding to 8,000 parameter vectors. For the higher resolution map with 5km×5km pixels, the number of parameter vectors increased to 9,000 when the number of maps decreased from $M = 2,000$ to $M = 1,000$ and lower. The same number of parameter vectors was required for 10km×10km and $M = 100$.

Figure 3 shows how the number of active pixels $|A|$, minimum and maximum ESS (over pixels) changed with each AMIS iteration for the eight configurations described previously. The AMIS framework produced a consistent efficiency when decreasing the number of active pixels and increasing the minimum ESS, with the exception of $M = 100$ maps. For this configuration both $|A|$ and $\min(ESS)$ had a slight lag in comparison to $M = 500$, $M = 1000$ and $M = 2000$. Although we remove pixels with $ESS \geq ESS^R$ from the active pixel sets, they are still benefiting from the additionally sampled parameters as can be seen from the $\max(ESS)$ dynamics.

Details of the map and the parameters sampled for resolution 5km×5km and $M = 2000$ are shown in Figure 4. Firstly, we plotted a map of mean prevalence for each pixel (fig.4 (a)), which ranged between 0.1 and 0.55. Second, a histogram of sampled prevalences, colour coded according to a propor-

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

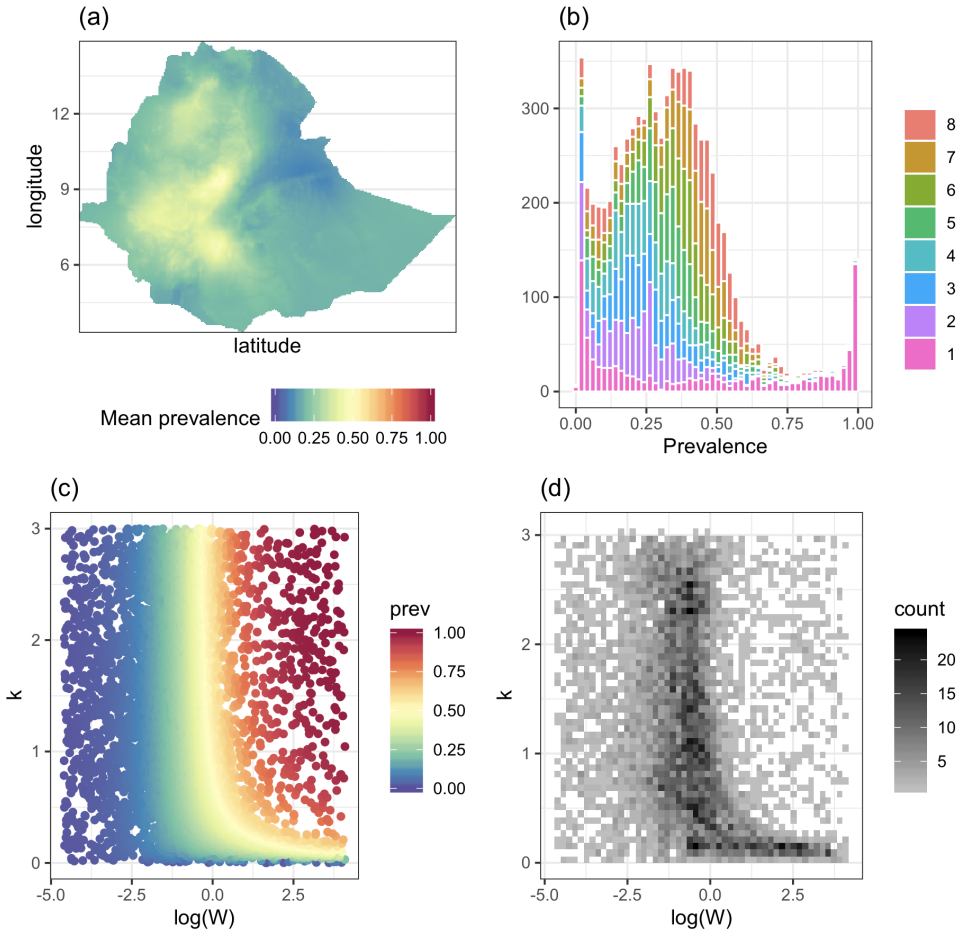


FIG 4. *Results for Ascaris lumbricoides in Ethiopia. (a) Map of mean prevalences constructed from survey data in Ethiopia (25). The resolution is 5km×5km and* $M = 2000$. *Sampled communities and prevalences are shown in Figure S1. (b) Histogram of sampled prevalences with colour coding according to the iteration of the AMIS algorithm. (c) Scatter plot of sampled model parameters and corresponding prevalence. (d) Density of sampled parameters. Prevalence and density distributions are shown as a function of log mean number of worms,* $\log(W)$, *and degree parasite aggregation,* k.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

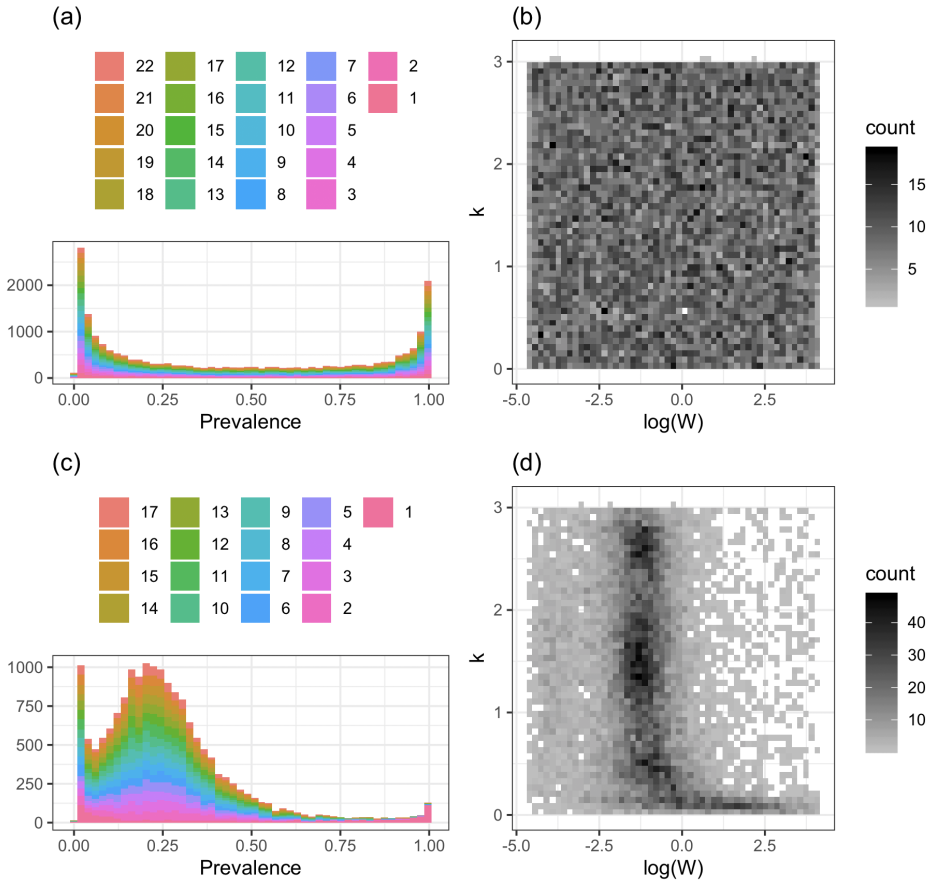


FIG 5. *Results for Ascaris lumbricoides in Ethiopia. At each AMIS iteration, the parameters were sampling from the prior (a)-(b); or using weights of all pixels (c)-(d). (a) and (c) shows histogram of sampled prevalences with colour coding according to the iteration of the AMIS algorithm. (c) and (d) shows scatter plot of sampled model parameters and corresponding prevalence. (d) Density of sampled parameters. Prevalence and density distributions are shown as a function of log mean number of worms,* $\log(W)$*, and degree parasite aggregation,* k*. The resolution is 5km*$\times$*5km and* $M = 2000$.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

tion sampled at each iteration, is shown in Fig.4(b). Here, iterations 2 and above targeted prevalences mostly between 0 and 0.5, as this is the range of mapped prevalences of *Ascaris* (Supplementary Figure S1). Third, Fig.4(c) shows prevalence as a function of log mean number of worms, $\log(W)$, and degree of parasite aggregation, k; this function has a complex "L-shaped" dependence with prevalences ranging from 0 to 1 as $\log(W)$ or k increase. Finally, when we look at the density of sampled parameters, the algorithm targeted areas which show the largest variability in terms of simulated prevalences, i.e along this "L-shaped" contour (fig.4 (d)).

The alternative methods we tested, i.e. based on sampling from the prior only, and based on AMIS with $|A| = M$ for all iterations, produced samples from different regions of parameter space (Figure 5). When sampling uniformly, we found that a histogram of posterior prevalences has a "U-shape", i.e. very low and very high values of prevalences are over-sampled, which does not correspond to the mapped prevalences (Figure SI). When sampling with the set of active pixels equal to all pixels, the algorithm targeted parameter regions similar to those of the adaptive AMIS, but still required twice as many iterations.

3.1.4. *Incorporating parameter dependencies.* A consistent relationship between mean worm burden and prevalence of *Ascaris lumbricoides* infection has been observed in a data coming from a range of geographically distinct human communities (34). Therefore, we have testing the performance of the AMIS framework when the prior of the parameters incorporates the dependence between mean worm burden, W, and degree of parasite aggregation, k. We have estimated the relationship between k and W using paired prevalence and mean intensity data from (34). This have led to the following prior for parameter: uniform prior for log of the mean number of worms $\log(W) \sim U[\log(0.01), \log(60)]$ and a Gaussian distribution for the degree of clumping $k \sim N(0.3337 + 0.0171 * W, \sigma(W))$ (Supplementary Figure S2). For details on fitting procedure see Appendix A in Supplement.

Figure 6 shows that nine iterations were required for the AMIS algorithm, which is comparable to eight iterations when applying AMIS without dependence between W and k (Fig. 4).

3.2. *Onchocerciasis in Togo.* Onchocerciasis is a filarial infection caused by *Onchocerca volvulus* and transmitted among humans via the bites of female *Simulium* blackflies. Onchocerciasis is responsible for skin disease, visual impairment including blindness, and excess mortality, which may be associated with epilepsy (35). Programmes for the control and elimination of the onchocerciasis have been targeted to the areas most affected and great

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

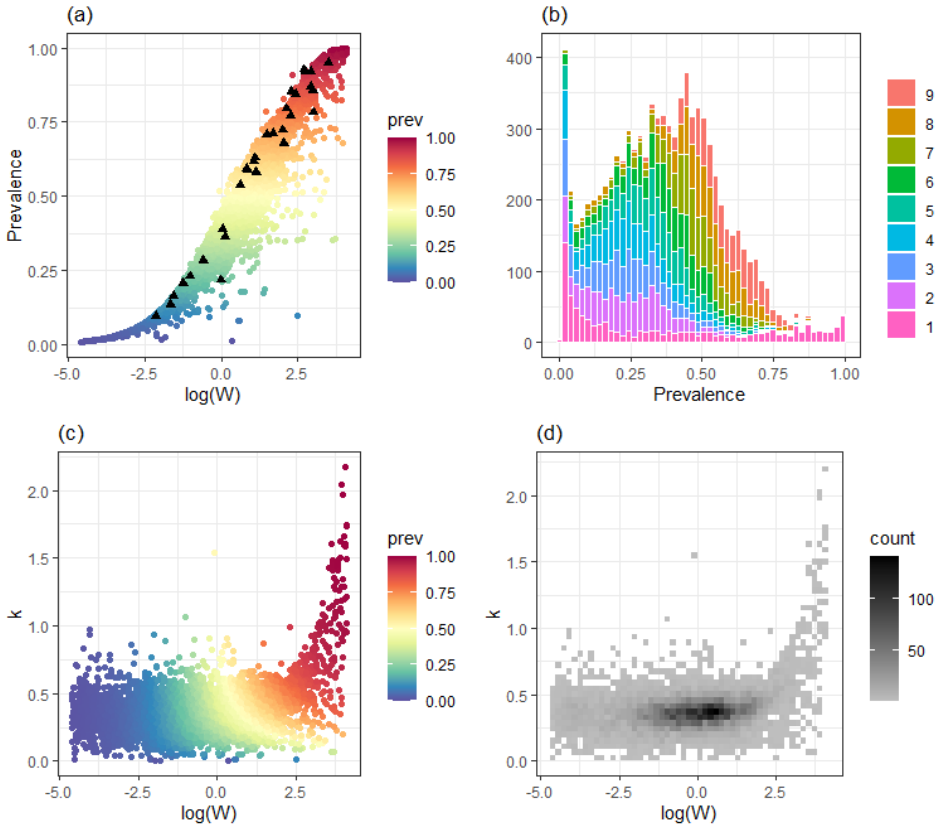


FIG 6. *Results for Ascaris lumbricoides in Ethiopia. (a) Sampled values of log mean number of worms, log(W) and corresponding values of prevalence (coloured dots) with the experimental data from (34) (black triangles). (b) Histogram of sampled prevalences with colour coding according to the iteration of the AMIS algorithm. (c) Scatter plot of sampled model parameters and corresponding prevalence. (d) Density of sampled parameters. Prevalence and density distributions are shown as a function of log mean number of worms,* $\log(W)$*, and degree parasite aggregation,* k*. The resolution is 5km×5km and* $M = 2000$.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

strides have been made, but challenges remain to achieve elimination of transmission (35). Some of these include the lack of efficacious drugs to kill the adult worms (which can live, on average, 10 years, but may live up to 15-20 years (36). Annual or bi-annual mass drug administration (MDA) with ivermectin is the main intervention strategy, but vector control (treating the vector breeding sites in fast flowing rivers) has also been used with success (37). Targeting vector control activities effectively would necessitate an in-depth knowledge of vector breeding site ecology, as well as a better understanding of hydrological conditions in rivers (38; 39; 40).

3.2.1. *Model and data.* We simulated onchocerciasis transmission in communities using the deterministic version of the EPIONCHO model (41). This model includes: age and sex structure of the human population; age- and sex-specific exposure to blackfly bites; dynamics of the mean number of fertile and non-fertile female worms per host; mean number of microfilariae per mg of skin in the human host and mean number of infective larvae per blackfly vector. All parameters, except the annual biting rate (the number of bites per person per year), abr, and k, the aggregation of adult warms, were set as in the supplementary material of (41). The annual biting rate determines the intensity of transmission and therefore the prevalence and intensity of infection.

We used maps of baseline (pre-control) microfilarial prevalence (by skin snip microscopy) for ages 5+ years at 5km$\times$5km spatial resolution developed in (4) for the area of the Onchocerciasis Control Programme in West Africa. The data consist of $M = 2,000$ samples for each pixel, with each individual prevalence map consisting of $I = 9,360$ pixels.

3.2.2. *Implementation of AMIS.* We set a uniform prior for the log of the annual biting rate: $\log(abr) \sim U[\log(100), \log(30,000)]$ and a uniform prior for the aggregation of adult warms $k \sim U[0, 3]$.

Regular distribution of ivermectin treatment in the form of MDA campaigns is the main current strategy for onchocerciasis control. (Although vector control was also implemented in Togo, we have focused on ivermectin MDA). As ivermectin is mostly microfilaricidal, and the microfilariae are responsible for morbidity and transmission to vectors, this type of intervention reduces disease progression in treated individuals and reduces transmission in the population. For each sampled parameter vector, we have run projections for 15 years under two levels of therapeutic coverage using the EPIONCHO model (41). In particular, we have simulated a coverage of 65% of the total population (the minimum coverage recommended by the World Health Organization), and an enhanced coverage of 80% (as an alterna-

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

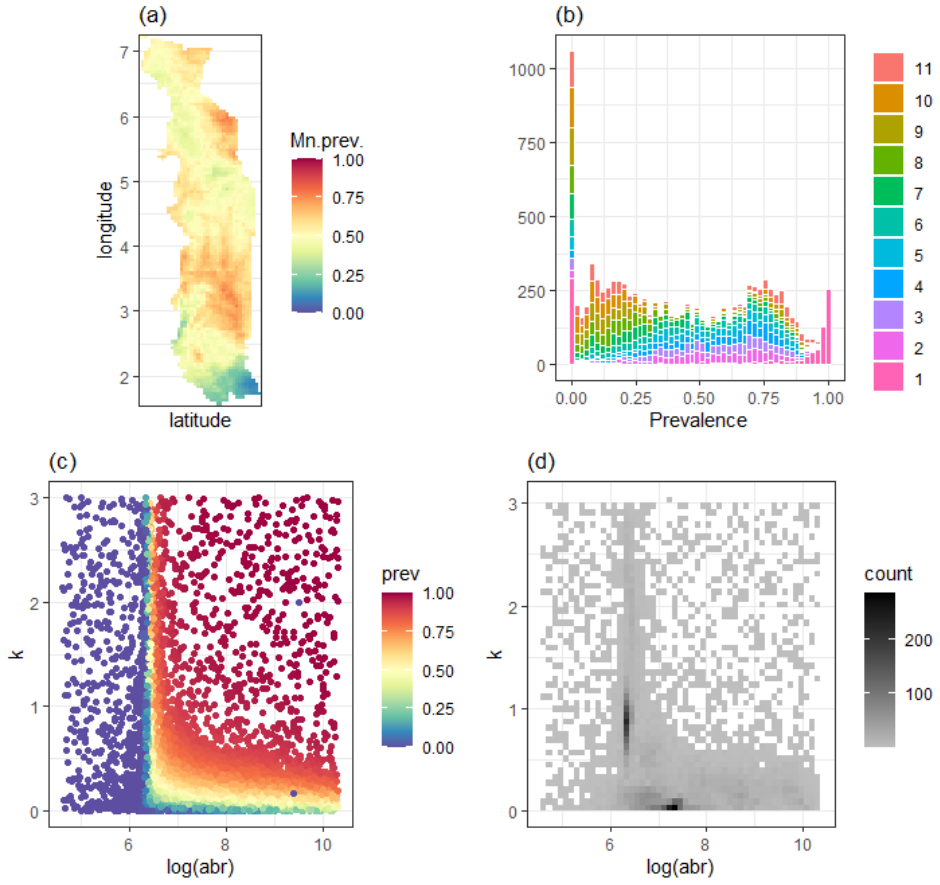


FIG 7. *Results for onchocerciasis in Togo. (a) Map of mean baseline prevalence (4). (b) Histogram of sampled prevalences with colour coding according to the various iterations of the AMIS algorithm. (c) Scatter plot of sampled model parameters and corresponding prevalence. (d) Density of sampled parameters. In (c) and (d), prevalence and density distributions are shown as a function of log annual biting rate,* $\log(abr)$*, and the aggregation of adult warms,* k*.*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

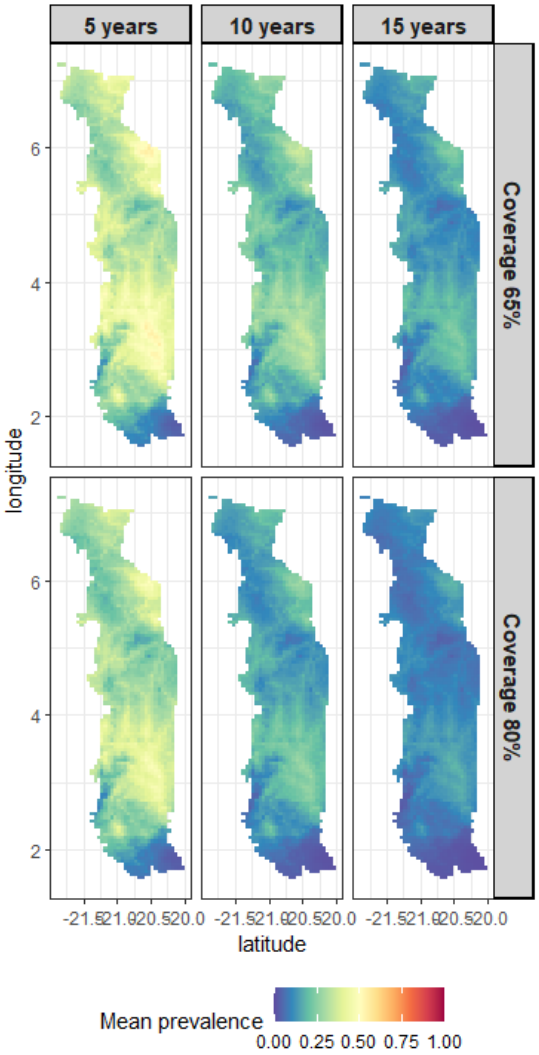


FIG 8. *Comparison of the impact of ivermectin coverage levels on elimination of onchocerciasis. The interventions compared are mass drug administration of ivermectin at 65% coverage, and mass drug administration at 80% coverage. Projections run using the EPIONCHO model (41).*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

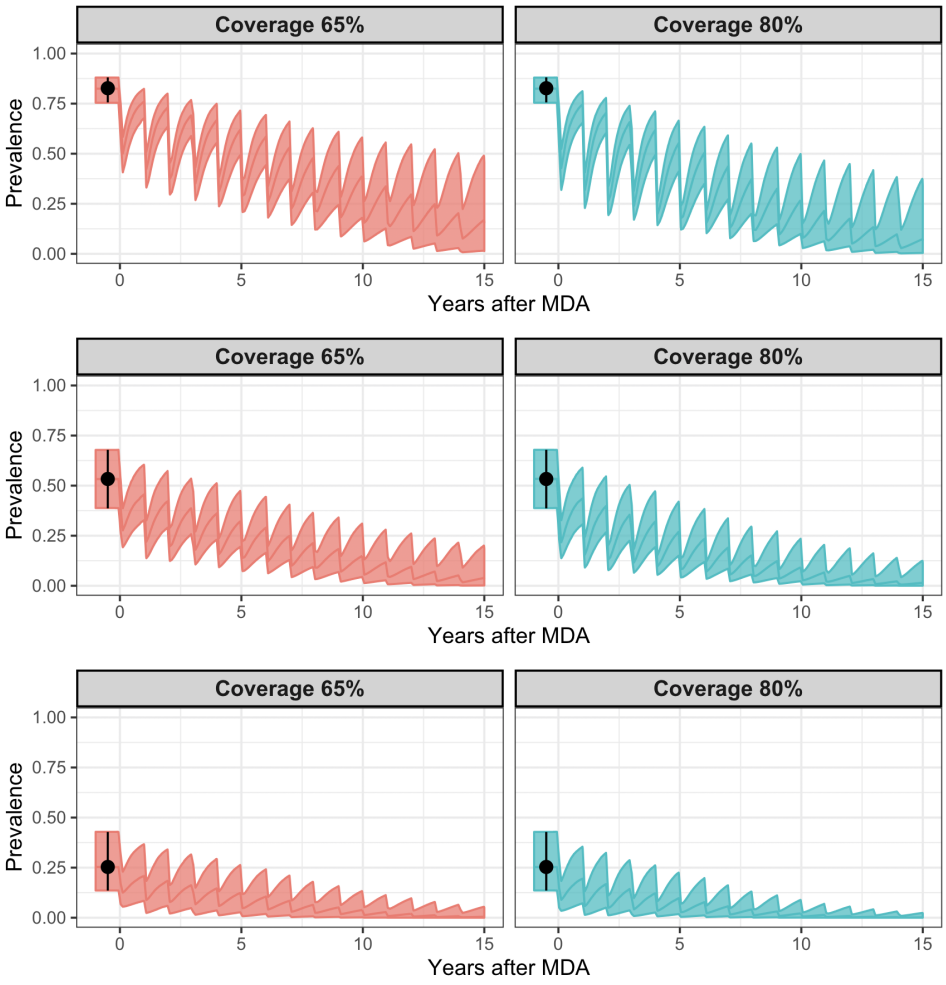


FIG 9. *Comparison of two strategies (minimum recommended 65% coverage of total population, and enhanced 80% coverage as an alternative treatment strategy) for onchocerciasis control and elimination at three levels of baseline endemicity (hyper-, meso and hypoendemicity). Projections were simulated using the EPIONCHO model . Projections were simulated using the EPIONCHO model (41). Mapped modelled prevalences are shown in black (4). Figures show mean and 95% inter-quantile range.*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

tive treatment strategy (42), the latter meaning that treatment is reaching nearly the entirety of the eligible population. In both cases, the proportion of systematic non-adherers was set at 5%.

3.2.3. *Results.* Similar to the case of ascarisis (Section 3.1), there is a strong heterogeneity in the spatial distribution of onchocerciasis prevalence in Togo (figure 7 (a)). The mean microfilarial prevalence can reach levels of up to 80% in holoendemic areas with high vector biting rates. Figures 7 (b)-(d) show sampled parameter vectors and simulated prevalences. The AMIS framework required 10 iterations to achieve $ESS \geq ESS^R$ in all pixels. Again, we can see that AMIS sampled most of the parameters along a contour resembling an "L-shape" with quite narrow regions corresponding to the largest variation of prevalences. A histogram of simulated prevalences has a high frequency for values close to zero and two peaks at around 0.25 and 0.8 (Fig.7 (b)).

We compared the impact of 15 years of annual ivermectin MDA with total population coverage 65% and 80%. Figure 8 shows reduction in microfilarial prevalence in endemic area 5, 10 and 15 years after start of MDA. Pre-control endemicity levels (i.e. pre-intervention infection prevalence and intensity) have been indicated as a crucial factor determining the success of intervention strategies to achieve elimination of transmission (42), with areas of higher baseline endemicity indicative of intense transmission (higher basic reproduction ratio) due to high blackfly vector density, where elimination is more difficult (in the absence of vector control). Furthermore, a modelling study using the stochastic version of the model, EPIONCHO-IBM, found that the resilience of the parasite population to MDA was markedly higher for lower levels of exposure heterogeneity k (43). Figure 9 shows the temporal dynamics of microfilarial prevalence under annual ivermectin MDA for three pixels corresponding to different levels of mean baseline microfilarial prevalence (0.8 for a hyperendemic/holoendemic situation; 0.5 for a mesoendemic setting, and 0.3 for a hypoendemic scenario). First, it can be seen that there is a good agreement between the mapped and simulated baseline prevalences. Second, uncertainty is propagated with time and simulations produce a wide range of projected prevalences after 15 annual rounds of MDA, especially for the hyperendemic scenario. This is due to equal sampling from different regions of the parameter space, where sets of $\{abr, k\}$ match similar levels of baseline prevalence but have their own characteristic resilience to MDA. Uncertainty can be reduced by utilizing post-control maps within the AMIS algorithm (see Section 3.4).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

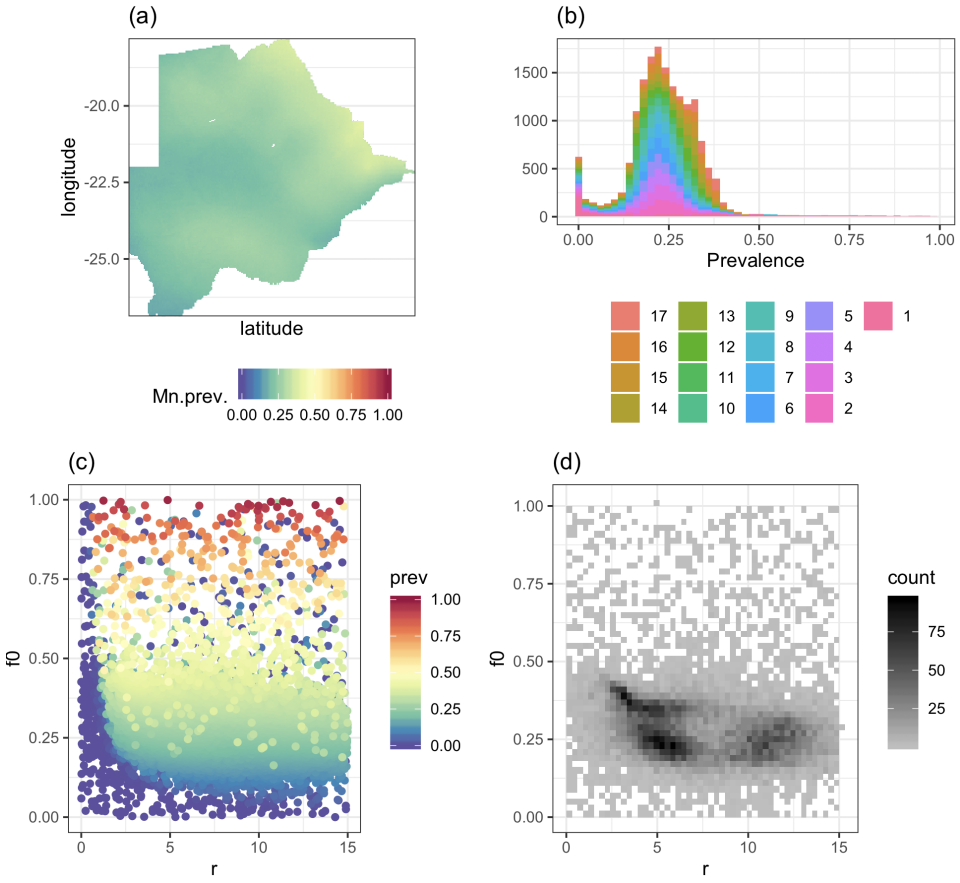


FIG 10. *Results for HIV in Botswana. (a) Map of mean prevalences in 2010 (7). The resolution is 5km×5km and* $M = 100$. *(b) Histogram of sampled prevalences with colour coding according to the various iterations of the AMIS algorithm. (c) Scatter plot of sampled model parameters and corresponding prevalence. (d) Density of sampled parameters. In (c) and (d), prevalence and density distributions are shown as a function of the rate of infection,* r, *and the fraction of the population in the at-risk category at the start of the epidemic,* f_0.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

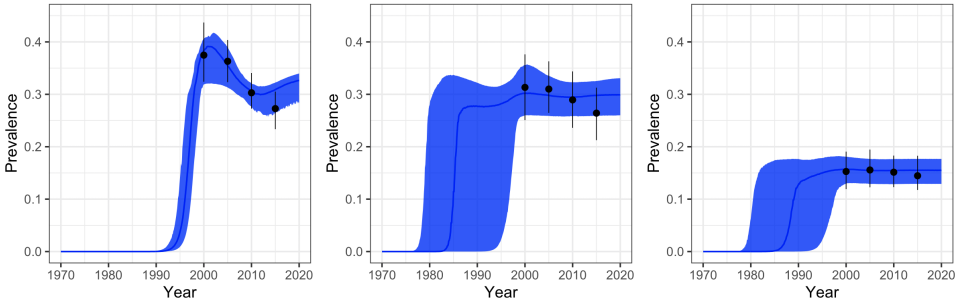


FIG 11. *Comparison of mapped and simulated prevalences at three levels of endemicity for HIV in Botswana. Mapped modelled prevalences are shown in black (7). Figures show mean and 95% inter-quantile range.*

3.3. *HIV infection in Botswana.* HIV/AIDS is a sexually transmitted infection with an estimated 6000 new infections occurring globally each day (44). It is a leading cause of disease burden in sub-Saharan Africa (7).

3.3.1. *Model and data.* We have used the model developed by the Joint United Nations Programme on HIV/AIDS (UNAIDS) called Estimation and Projection Package (EPP) (45). This is a deterministic Susceptible-Infected model with the population divided into three groups: a not-at-risk group, an at-risk group, and an infected group. Prevalence dynamics are defined by four parameters: the start time of the epidemic, t_0; the fraction of the population in the at-risk category at the start of the epidemic, f_0; the rate of infection, r (governs the rate at which people in the susceptible population become infected), and the behavioral response, ϕ (this parameter affects the distribution of new entrants to the not at-risk and at-risk categories). We set the initial size of the population to 10,000. For full description of the model see (46).

We have used maps of estimated prevalence of HIV among adults (aged 15–49 years) in sub-Saharan Africa from 2000 to 2017 (7). These maps were at 5km×5km spatial resolution. We have extracted maps for Botswana, which had $I = 29,174$ pixels per map. However, only maps for mean, lower bound (2.5th percentile) and upper bound (97.5th percentile) were avail-

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

able, which were calculated from the 1,000 posterior draws (7). As this case study is an illustrative example to test how the AMIS framework performs when the dimension of the parameter space gets bigger, we have used these values to re-populate 100 individual maps by assuming that mapped prevalences were drawn from two half-normal distributions with means equal to the mapped mean, and standard deviations as a function of upper/lower bound.

3.3.2. *Implementation of AMIS.* We set a uniform prior for the rate of infection: $r \sim U[0, 15]$, a uniform prior for the fraction of the population in the at-risk category at the start of the epidemic $f_0 \sim U[0, 1]$, a uniform prior for the start time of the epidemic $t_0 \sim U[1970, 1990]$, and a uniform prior for the behavioral response: $\phi \sim U[-200, 400]$. Here priors for r, f_0 and t_0 are similar to those used in (46), excepting the prior for ϕ, which was set as a normal distribution with mean 0.

We have chosen to use maps of estimated prevalence of HIV in 2000, 2005, 2010 and 2015 (7). Therefor we have $K = 4$ time points to match for each simulated epidemic curve.

3.3.3. *Results.* The mean mapped prevalence distribution of HIV among adults in Botswana was more homogeneous in comparison to that of ascariasis (Section 3.1) or onchocerciasis (Section 3.2)) (which is not surprising given the exacting ecological requirements for these two helminth infections). HIV prevalence ranged between 0.151 and 0.401 in 2010 (Fig.10(a) (7)). Similar ranges were observed for other years as well: 0.155-0.371 in 2005; 0.151-0.332 in 2010, and 0.137-0.318 in 2015.

Figures 10(b)-(d) show the results of the AMIS framework. The AMIS framework required 17 iterations, i.e., 17,000 parameter vectors (Fig. 10(b)). The algorithm targeted areas of parameter space corresponding to prevalences between 0.1 and 0.4, which is the range of mapped prevalences in 2000. The figure 10 shows a scatter plot and marginal density distribution for two of four parameters: the rate of infection, and the fraction of the population in the at-risk category at the start of the epidemic. Scatter plots for all parameters are shown in supplementary Figure S2. It can be seen that the relationship between parameters and infection prevalence is more complex than in the previous two case.

We have found high uncertainty about prevalence dynamics before the time of mapped data, i.e. year 2000 (Figure 11). The authors in (46) have restricted prevalence in 1980 to be smaller than 10%. This restriction would be straightforward to implement in our algorithm by introducing an additional map in 1980 with values of prevalence $q_{i,m} \sim U[0, 0.1]$.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

3.4. *Malaria in the Democratic Republic of the Congo.* Malaria is a serious vector-born disease caused by the parasitic protozoan *Plasmodium* and transmitted among humans via the bites of female *Anopheles* mosquitoes (47). Although malaria infection prevalences have decreased by half in the past decade due to global efforts, there are still an estimated half a million of deaths a year (48). Mathematical modelling of malaria transmission can help to identify localised areas where control effectiveness is failing and provide a rational for planning intervention programmes (49).

3.4.1. *Model and data.* We have used the OpenMalaria platform to simulate Plasmodium falciparum parasite prevalence (PfPR) under a microscopy-based parasite detection (50). The OpenMalaria platform combines an individual-based stochastic simulation model of malaria in humans with a deterministic model of malaria in mosquitoes. All parameters, except the annual Entomological Inoculation Rate (the number of infectious mosquito bites per person per year), $aEIR$, and the variance between individuals in parasite densities (of asexual stages in erythrocytes), σ_i^2, were set as in (51).

Data from PfPR surveys in Africa in 2000 were used to construct a map of PfPR prevalence for children aged 2-10 years at 5km×5km spatial resolution in sub-Saharan Africa (1). We have extracted maps for the DRC, which had $I = 108,662$ pixels for each map. The data has 100 sample draws of prevalences for each pixel. This case study was different from the onchocerciasis case as we have a large number of pixels, but a small number of maps for each individual pixel.

3.4.2. *Implementation of AMIS.* We set a uniform prior for the log of annual Entomological Inoculation Rate: $\log(aEIR) \sim U[\log(1), \log(500)]$, and a uniform prior for variance in parasite density among individuals, $\sigma_{i,2} \sim U[0.1, 10]$. We set population size to 10,000 and run the model for 100 years to get to an epidemic equilibrium. For details on model simulation see (50).

We chose to use maps of the estimated prevalence of malaria in 2005 and 2010. These years were chosen arbitrary, as we were interested to use $K = 2$ time points to match for each simulated epidemic curve. Furthermore, we simulated projections of prevalence up to year 2015 to be able to compare the predicted and mapped prevalences.

In order to simulate malaria transmission from 2000 to 2015, we had to account for the implementation of interventions during these years. Insecticide-treated bednet (ITN) coverage in the DCR in 2015 was downloaded from the *malariaAtlas* Project (MAP) (52). We ignored the coverage of ACTs as this was considered to be low in the DRC ($< 2\%$ in 2000-2010 and $< 16\%$ in 2015 (1)). We assumed spatial homogeneity in ITN coverage, which was

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

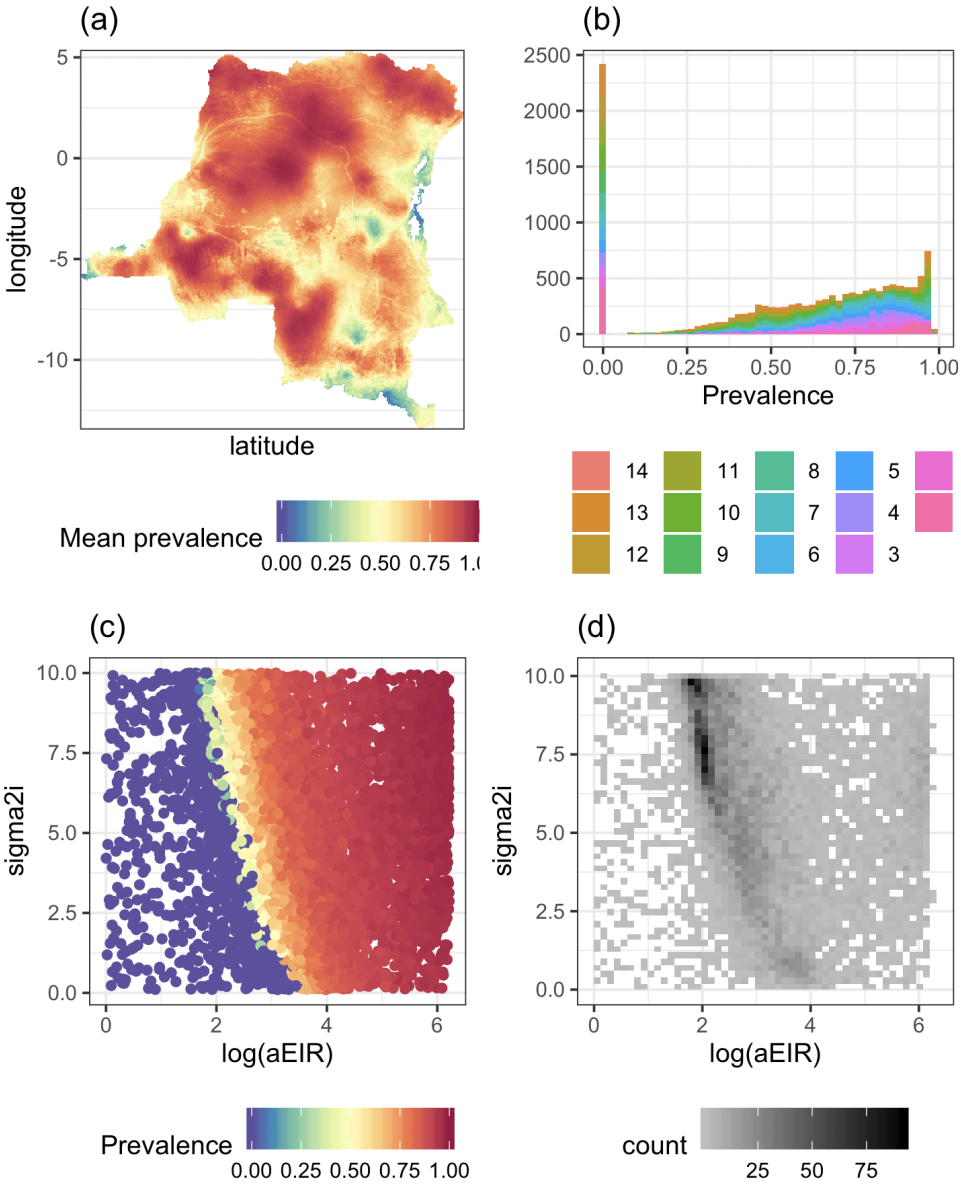


FIG 12. *Results for malaria in the Democratic Republic of the Congo. (a) Map of mean prevalences in 2005 (1). The resolution is 5km×5km and $M = 100$. (b) Histogram of sampled prevalences with colour coding according to the various iterations of the AMIS algorithm. (c) Scatter plot of sampled model parameters and corresponding prevalence. (d) Density of sampled parameters. In (c) and (d), prevalence and density distributions are shown as a function of log annual Entomological Inoculation Rate, aEIR, and the variance between individuals in parasite densities, σ_i^2. Projections were simulated using the OpenMalaria model (51).*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

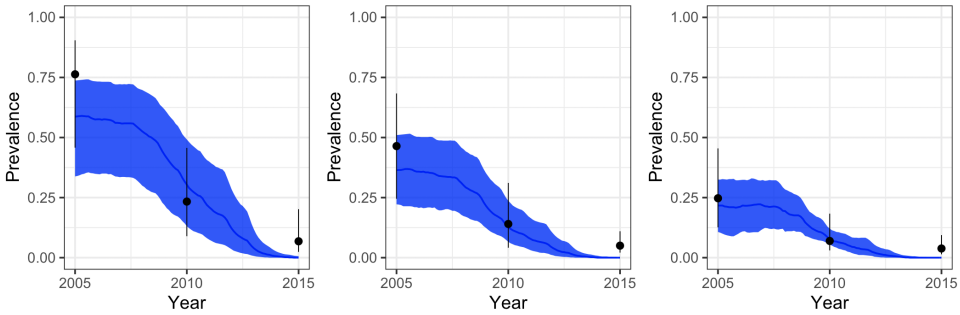


FIG 13. *Comparison of mapped and simulated prevalences at three levels of endemicity for Plasmodium falciparum malaria in the DRC. Mapped prevalences are shown in black (1). Projections were simulated using the OpenMalaria model (50). Figures show mean and 95% inter-quantile range.*

set to a mean 76%. To calculated annual coverage of ITNs, we used data on total ITNs distributed in the DRC in 2005-2014 (53), and assumed that the average mapped coverage of 76% was achieved by year 2015 (for details see Supplementary section B).

3.4.3. *Results.* The corresponding mean prevalence map and the results of AMIS parameter estimation are shown in Figure 12. It can be seen from fig. 12(a) that the mean prevalence of malaria in children had a large range of values, i.e. from 0.2 to 0.9. Furthermore, pixels with high/low mean prevalence are spatially clustered. Therefore, it would be expected that many pixels should have very similar prevalence distribution profiles, which would be beneficial for implementing the AMIS algorithm across all pixels of the map.

We found that prevalence exhibits a nonlinear relationship with respect to $aEIR$ and variabilityvariance in parasite densities among individuals, σ_i^2, exhibiting a similar tendency to that in ascariasis (Section 3.1) and onchocerciasis case (Section 3.2) (fig. 12(c)). The AMIS framework required 14 iterations to achieve $ESS \geq ESS^R$ (fig. 12(d)). As can be see from the plot of densities of sampled parameters, AMIS targeted the region of the parameter space corresponding to the largest variation in prevalence (fig.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

12(d)).

Figure 13 shows a comparison of mapped and simulated prevalences at three levels of malaria endemicity. Our results demonstrate a sunstantial decline in PfPR prevalence for children aged 2-10 years in 2015, which is over-estimated in comparison to the mapped prevalences. On average, the simulated epidemic curves tend to be lower than mapped prevalences in 2005 as well, which is in contrast to the onchocerciasis case, where simulated baseline prevalences were similar to the mapped prevalences (Fig. 9).

4. Discussion. Our proposed framework allows the distribution of prevalences from high-resolution maps to be projected forward into the future under the transmission dynamics of complex disease transmission models. Below we discuss the benefits and limitations of the methodology.

4.1. *Parameter estimation based on prevalences at a national level.* The diversity of diseases studied and the range of mathematical models used, allowed us to explore the performance of our framework under different conditions. As the results in the previous section showed, the AMIS framework performed well for all four case studies. The proposed algorithm, utilizing the combination of all the samples from multiple proposals, was suitable for the task of parameter estimation based on geospatial maps, as it allowed us to explore complex dependencies between parameters and modelled prevalences with relatively small numbers of transmission model simulations. An alternative strategy where parameters are sampled for each individual pixel independently would require much higher numbers of simulations, as the number of pixels in a map can be large. We used the same simulations for each pixel, with extra computational time required to calculate weights and ESS at each iteration.

4.2. *Using weights of active pixels.* The algorithm we propose has an additional step of calculating the average weight of parameter vectors based on individual weights of active pixels. In section 3.1 we compared this approach with two alternative methods, namely: based on sampling from the prior only, and based on AMIS with $|A| = M$ for all iterations. As can be seen from Table 1, the method based on weights of active pixels required 2-3 times fewer iterations then the algorithm based on sampling from the prior only or based on AMIS without removing pixels from the active set. The latter algorithm targets simulations towards the pixels that have already been well served by earlier iterations of the algorithm, and not towards the required posterior prevalences for all of the pixels, for example simulated prevalences between 0.5 and 0.75 (Fig. 5).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

4.3. *Influence of size of geospatial maps.* In Section 3.1 we found that the number of pixels (i.e. $I = 11,369$ pixels and $I = 37,695$ pixels for the prevalence of *Ascaris* in school-aged children in Ethiopia) had little influence on AMIS performance in terms of the number of iterations required (Table 1). The number of individual prevalence maps used for parameter sampling also had a minor influence, with the number of AMIS iterations equal to 8 or 9 when the number of maps varied from 100 to 2000. A slightly lower increase in minimal ESS was observed for $M = 100$, but this did not translate into any additional iterations. This suggests that having a higher number of maps will not require more transmission model simulations, but in fact can decrease this number in comparison to a smaller number of maps. Having a smaller number of maps for each pixel can make it harder to match the sampled prevalences to the distribution of observed prevalences which translates into extra iterations. This may be because there is a spatial correlation between neighbouring pixels, so simulated prevalences close to some observed prevalences at particular pixels will be similar to those in neighbour pixels.

4.4. *Setting priors of parameters.* The AMIS framework applied to transmission models involves defining a prior distribution for model parameters. In most case studies, we used a uniform distribution with parameter ranges informed by the available literature on biological processes governing infection transmission or previous analysis of mathematical models. Our framework allows incorporating specific knowledge of the parameter space, for example excluding particular areas/combinations of parameters that have been deemed biologically unfeasible, or introducing dependencies between parameters and can be informed by additional data. Practical application of this has been demonstrated for *A.lumbricoides* in section 3.1.4, where parameters for the degree of parasite aggregation were sampled from the Gaussian distribution conditioned on the values of mean number of warms.

4.5. *Impact of intervention programs.* In section 3.2 we have sampled parameters for onchocerciasis in Togo, simulated the transmission model further forward in time under two levels of coverage and applied the weights to obtain the spatial distribution of the projections. We found a good agreement between the mapped and simulated baseline prevalences. However, the uncertainty was propagated with time and simulations produce a wide range of future projected prevalences. Uncertainty can be reduced by utilizing post-control maps within the AMIS algorithm.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

4.6. *Comparison to Bayesian melding.* A probabilistic approach, called Bayesian melding, combines via logarithmic pooling two priors: one implicit and one explicit, on each input and output (54). Bayesian melding has been proposed and used to account for uncertainty in parameters and model projections for HIV (46). The initial stage of the algorithm required 200,000 combinations of the input parameters from their prior distribution and produced 373 unique epidemic curves fitted to aggregated data on HIV prevalence in urban clinics (46). This data would be equivalent to a single pixel in our case. As shown in Section 3.3, the AMIS algorithm offers improved computation efficiency and higher resolution in comparison to this application of Bayesian melding.

4.7. *Using post-control maps.* For the malaria case study in Section 3.4, our results demonstrate that when sampling parameters using baseline as well as post-control maps, care has to be taken that the simulated epidemiological curves are able to support mapped dynamics. We assumed that 15% coverage of ITNs was achieved by 2010 (Supplementary table S1), but this coverage would not reduce mean prevalence from 0.75 to 0.25, as seen in Figure 13. The AMIS framework produced a trade-off between fitting to maps in 2005 and 2010. However, the schedule of ITNs coverages used within the model was better suitable for pixels with low baseline endemicity. Therefore pixels should be divided into regions with different histories of control. The AMIS framework can then be applied within each region.

4.8. *Tuning of hyper-parameters.* The proposal distribution adapts to the target by locally fitting a mixture component to areas which correspond to pixels with low effective sample sizes. Therefore the performance of the algorithm depends on the choice of threshold ESS^R. In our case studies we have set $ESS^R = 2000$. The cost of AMIS sampling has the following components: fitting a mixture of distributions; sampling parameters from the mixture, performing model simulations for a parameter vector and calculating values of ESS. When the latter two are computationally expensive, lower values of ESS^R would be advisable.

We have used a prior for sampling just for a single iteration. In (16), the authors set half of the iterations to be sampled from the prior. In our case, we don't know in advance how many iterations will be required to have enough parameter sets to satisfy the condition $min(ESS) \geq ESS^R$. However, our results for all case studies indicate that having sampled 1000 parameter sets was enough to gauge important regions of parameter space for further analysis, at least in 2 to 4 dimensions. We also had the number of samples per iteration fixed to $N_t = 1000$. We anticipate that decreasing the value

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

of N_t could lead to a lower number of parameter sets required to achieve $\min(ESS) \geq ESS^R$, but at a cost of increasing the number of iterations, which will add an additional computational cost due to the requirement to recalculate the weights and ESS after each iteration.

The stability of the algorithm requires that the parameter space has been sufficiently explored in the initial iteration. Low values of ESS would show that there are many parameter samples which carry relatively low weight. However ESS can be deceptively high when all the simulations fall in regions with equally low weight. We avoid this by ensuring that the initial iteration comprises sufficiently many samples so that at least some of them fall in regions with high weight. Depending on mapped prevalences and model behaviour, the distribution of ESS can have a range of values after the first and subsequent iterations. For example in the case of *Ascaris*, we had that $\min(ESS) = 67.4$ and $\max(ESS) = 435.1$ after the first iteration, which are high in comparison to a number of sampled parameters, i.e. 1000 (Fig. 3).

4.9. *Alternative AMIS procedures.* Recently, a modified AMIS has been proposed, where an importance sampling distribution at iteration t is built based on samples from iteration $t-1$ but weights for all iterations are recalculated after the last iteration (55). A simpler recycling strategy could offer computational savings, but this would lose the advantage of the adaptive nature of the AMIS framework we proposed, i.e. to utilize information on the ESS based on all sampled parameter vectors rather then only the subset from the latest iteration.

A further class of adaptive importance samplers has been proposed, in which the adaptation is driven by independent parallel or interacting Markov chain Monte Carlo chains (56). Parallelisation might be a promising route, for example, by sampling N_t/d parameter vectors and running the model and projections on d parallel units, then combining them into a single iteration and using aggregated parameter vectors to sample next N_t/d parameter vectors.

5. Conclusions. Infectious diseases remain an important health problem worldwide. We have introduced a novel method to sample parameters of disease transmission models given high-resolution prevalence maps. Our strategy of using a mixture of Gaussian distributions based on a transmission model's similarity to mapped prevalence distributions leads to an efficient exploration of the state space. We have extended the methodology to include maps for multiple time points. In future work, applications of the methodology that account for routine surveillance data would allow greater

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

epidemiological insight into providing tools for policy at a local level, bringing infectious diseases under control, or even setting out the pathway for elimination of transmission.

Acknowledgments. The authors gratefully acknowledge funding of the NTD Modeling Consortium by the Bill and Melinda Gates Foundation [OPP1152057, OPP1053230, OPP1156227]. The views, opinions, assumptions or any other information set out in this article should not be attributed to the Bill and Melinda Gates Foundation or any person connected with the Bill and Melinda Gates Foundation.

SUPPLEMENTARY MATERIAL

Supplement: Integrating geostatistical maps and transmission models using adaptive multiple importance sampling
(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide additional supporting plots for *Ascaris* and HIV.

The code is available at https://github.com/rretkute/AMIS$_I D$.

References.

[1] Bhatt S., Weiss D.J., Cameron E., BisanzioD., Mappin B., Dalrymple U.,1 Battle K., Moyes C.L., Henry A., Eckhoff P.A., Wenger E.A., Brit O., Penny M.A., Smith T.A., Bennett A., Yukich J., Eisele T.P., Griffin J.T., Fergus C.A., Lynch M., Lindgren F., Cohen J.M., Murray C.L.J., Smith D.L., Hay S.I., Cibulskis R.E., Gething P.W. (2015) The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015 *Nature* **526** 207–211.

[2] Karagiannis-Voules D.A., Biedermann P., Ekpo U.F., Garba A., Langer E., Mathieu E., Midzi N., Mwinzi P., Polderman A.M., Raso G., Sacko M., Talla I., Tchuent L.A., Tour S., Winkler M.S., Utzinger J., Vounatsou P. (2015) Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: a systematic review and geostatistical meta-analysis. *Lancet Infect Dis.* **15** 74-84.

[3] Pigott D.M., Bhatt S., Golding N., Duda K.A., Battle K.E., Brady O.J., Messina J.P., Balard Y., Bastien P., Pratlong F., Brownstein J.S., Freifeld C.C., Mekaru S.R., Gething P.W., George D.B., Myers M.F., Reithinger R., Hay S.I. (2014) Global distribution maps of the leishmaniases. *Elife* **3** e02851.

[4] O'Hanlon, SJ ; Slater, HC; Cheke, RA; Boatin, BA; Coffeng, LE; Pion, SDS; Boussinesq, M; Zoure, HGM; Stolk, WA; Basanez, MG. (2016) Model-Based Geostatistical Mapping of the Prevalence of Onchocerca volvulus in West Africa. *PLoS Negl Trop Dis.* **10** e0004328.

[5] Bhatt S., Gething P.W., Brady O.J., Messina J.P., Farlow A.W., Moyes C.L., Drake J.M., Brownstein J.S., Hoen A.G., Sankoh O., Myers M.F., George D.B., Jaenisch T., Wint G.R., Simmons C.P., Scott T.W., Farrar J.J., Hay S.I. (2013) The global distribution and burden of dengue *Nature* **496** 504–507.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[6] Simarro P.P., Cecchi G., Franco J.R., Paone M., Diarra A., Ruiz-Postigo J.A., Fvre E.M,, Mattioli R.C., Jannin J.G. (2012) Estimating and mapping the population at risk of sleeping sickness. *PLoS Negl Trop Dis.* **6** e1859.

[7] Dwyer-Lindgren L., Cork M.A., Sligar A., Steuben K.M., Wilson K.F., Provost N.R., Mayala B.K., VanderHeide J.D., Collison M.L., Hall J.B., Biehl M.H., Carter A., Frank T., Douwes-Schultz D., Burstein R., Casey D.C., Deshpande A., Earl L., El Bcheraoui C., Farag T.H., Henry N.J., Kinyoki D., Marczak L.B., Nixon M.R., Osgood-Zimmerman A., Pigott D., Reiner R.C. Jr, Ross J.M., Schaeffer L.E., Smith D.L., Davis Weaver N., Wiens K.E., Eaton J.W., Justman J.E., Opio A., Sartorius B., Tanser F., Wabiri N., Piot P., Murray C.J.L., Hay S.I. (2019) Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* **570** 189-193.

[8] Mosser J.F., Gagne-Maynard W., Rao P.C., Osgood-Zimmerman A., Fullman N., Graetz N., Burstein R., Updike R.L., Liu P.Y., Ray S.E., Earl L., Deshpande A., Casey D.C., Dwyer-Lindgren L., Cromwell E.A., Pigott D.M., Shearer F.M., Larson H.J., Weiss D.J., Bhatt S., Gething P.W., Murray C.J.L., Lim S.S., Reiner R.C. Jr, Hay S.I. (2019) Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000-2016: a spatial and temporal modelling study. *Lancet* **4** 393 1843–1855.

[9] Hay S.I., Battle K.E., Pigott D.M., Smith D.L., Moyes C.L., Bhatt S., Brownstein J.S., Collier N., Myers M.F., George D.B., Gething P.W. (2013) Global mapping of infectious disease *Philos Trans R Soc Lond B Biol Sci.* **368** 20120250

[10] Giorgi E., Diggle P.J., Snow R.W., Noor A.M. (2018). Geostatistical Methods for Disease Mapping and Visualisation Using Data from Spatio-temporally Referenced Prevalence Surveys. *International Statistical Review* **86** 571–597.

[11] Amoah B., Diggle P.J., Giorgi E. (2020) A Geostatistical Framework for Combining Spatially Referenced Disease Prevalence Data from Multiple Diagnostics. *Biometrics* **76** 158–170.

[12] Hollingsworth T.D. (2018) Counting Down the 2020 Goals for 9 Neglected Tropical Diseases: What Have We Learned From Quantitative Analysis and Transmission Modeling? *Clin Infect Dis.* **66** (suppl4) S237–S244.

[13] Cornuet J.M., Marin J.M., Mira A., Robert C.P. (2012) Adaptive Multiple Importance Sampling. *SCANDINAVIAN JOURNAL OF STATISTICS* **39** 798–812.

[14] Ripley B.D. (1987) *Stochastic simulation.* John Wiley and Sons Inc., New York.

[15] Veach E., Guibas L. (1995) Optimally Combining Sampling Techniques for Monte-Carlo Rendering. in *Proceedings of the 22nd annual conference on Computer graphicsand interactive techniques*, SIGGRAPH **95** 419–428.

[16] Siren J., Marttinen P. Corander J. (2010). Reconstructing population histories from single-nucleotide polymorphism data. *Mol.Biol.Evol.* **28** 673–683.

[17] Sbert M., Havran V. (2017) Adaptive multiple importance sampling for general functions. *The Visual Computer* **33** 6–8.

[18] Elvira V., Santamaria I. (2019) Multiple Importance Sampling for Efficient Symbol Error Rate Estimation. *IEEE SIGNAL PROCESSING LETTERS* **2** 420–424.

[19] Anderson R.A., May R.M. (1992) *Infectious Diseases of Humans*, 1st ed. Oxford Univ. Press, Oxford.

[20] Keeling M.J, Rohani P. (2008). *Modeling Infectious Diseases in humans and animals*, 1st ed. Princeton University Press, New Jersey.

[21] Touloupou P., Retkute R., Hollingsworth T.D., Spencer S.E.F. (2020) Statistical methods for linking geostatistical maps and transmission models: Application

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

to lymphatic filariasis in East Africa. Submitted to *Spatial and Spatio-temporal Epidemiology*.

[22] Kish L. (1965). *Survey Sampling*. 1st ed. Wiley, New York.

[23] Cappe O., Douc R., Guillin A., Marin J.M., Robert C.P. (2008) Adaptive importance sampling in general mixture classes. *Statistics and Computing* **18** 447–459.

[24] Scrucca L., Fop M., Murphy T.B., Raftery A.E. (2016) mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8** 289–317.

[25] Pullan R.L., Smith J.L., Jasrasaria R., Brooker S.J. (2014) Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasit Vectors. BioMed Central* **7** 37.

[26] Ntamabyaliro N.Y., Burri C., Nzolo D.B., Engo A.B., Lula Y.N., Mampunza S.M., Nsibu C.N., Mesia G.K., Kayembe J.M.N., Likwela J.L., Kintaudi L.M., Tona G.L. (2018) Drug use in the management of uncomplicated malaria in public health facilities in the Democratic Republic of the Congo. *Malar J* **17** 189.

[27] Norhayati M., Fatmah M.S., Yusof S., Edariah A.B. (2003) Intestinal parasitic infections in man: a review. *Med J Malaysia* **58** 296–305.

[28] Anderson RM, May RM (1985) Helminth infections of humans: mathematical models, population dynamics, and control. *Advances in Parasitology* **24** 1–101.

[29] Rue H., Martino S., Chopin N. (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B.* **71** 319–392.

[30] Lindgren F., Rue H., Lindstrom J. (2011) An explicit link between Gaussian fields and Gaussian Markov random elds: The SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B.* **73** 423–498.

[31] Brooker S., Clements A.C.A., Bundy D.A.P. (2006) Global Epidemiology, Ecology and Control of Soil-Transmitted Helminth Infections. *Advances in Parasitology* **62** 221–261.

[32] Hijmans, RJ (2019) Raster: Geographic Data Analysis and Modeling.

[33] Fowler, AC; Hollingsworth, TD. (2016) The Dynamics of Ascaris lumbricoides Infections. *Bull Math Biol.* **78** 815–833.

[34] Guyatt H.L., Bundy D.A., Medley G.F., Grenfell B.T (1990) The relationship between the frequency distribution of Ascaris lumbricoides and the prevalence and intensity of infection in human communities. *Parasitology* **101** 139–43.

[35] Colebunders R., Basanez M.G, Siling K., Post R.J., Rotsaert A., Mmbando B., Suykerbuyk P., Hopkins A. (2018) From river blindness control to elimination: bridge over troubled water. *Infect. Dis Poverty* **7** 21.

[36] Plaisier A.P., van Oortmarssen G.J., Remme J., Habbema J.P. (1991). The reproductive lifespan of *Onchocerca volvulus* in West African savanna. *Acta Trop.* **48** 271–284.

[37] Hougard J.M., Alley E.S., Yameogo L., Dadzie K.Y., Boatin B.A. (2001). Eliminating onchocerciasis after 14 years of vector control: A proved strategy. *J. Infect. Dis.* **184** 497–503.

[38] Cheke R.A., Basanez M.G., Perry M., White M.T., Garms R., Obuobie E., Lamberton P.H., Young S., Osei-Atweneboana M.Y., Intsiful J., Shen M., Boakye D.A., Wilson M.D. (2015) Potential effects of warmer worms and vectors on onchocerciasis transmission in West Africa. *Philos Trans R Soc Lond B Biol Sci* **370** 20130559.

[39] Routledge I., Walker M., Cheke R.A., Bhatt S., Baleguel Nkot P.,

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Matthews G.A., Baleguel D., Dobson, Wiles T.L., Basanez M.G. (2018). Modelling the impact of larviciding on the population dynamics and biting rates of Simulium damnosum (s.l.): implications for vector control as a complementary strategy for onchocerciasis elimination in Africa. *Parasit Vectors* **11** 316.

[40] Jacob B.G., Loum D., Lakwo T.L., Katholi C.R., Habomugisha P., Byamukama E., Tukahebwa E., Cupp E.W., Unnasch T.R. (2018). Community-directed vector control to supplement mass drug distribution for onchocerciasis elimination in the Madi mid-North focus of Northern Uganda. *PLoS Negl Trop Dis* **12** e0006702.

[41] Walker, M; Stolk, WA; Dixon, MA; Bottomley, C; Diawara, L; Traor, MO; de Vlas, SJ; Basez MG. (2017) Modelling the elimination of river blindness using long-term epidemiological and programmatic data from Mali and Senegal. *Epidemics* **18** 4–15.

[42] Verver S., Walker M., Kim Y.E., Fobi G., Tekle A.H., Zoure H.G.M., Wanji S., Boakye D.A., Kuesel A.C., de Vlas S.J., Boussinesq M., Basanez M.G., Stolk W.A. (2018) How Can Onchocerciasis Elimination in Africa Be Accelerated? Modeling the Impact of Increased Ivermectin Treatment Frequency and Complementary Vector Control. *Clin Infect Dis.* **66**(suppl4) S267–S274.

[43] Hamley J.I.D, Milton P., Walker M., Basez M.G. (2019) Modelling exposure heterogeneity and density dependence in onchocerciasis using a novel individual-based transmission model, EPIONCHO-IBM: Implications for elimination and data needs. *PLoS Negl Trop Dis.*

[44] Kharsany A.B.M., Karim Q.A. (2016) HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. *Open AIDS J.* **10** 34–48.

[45] UNAIDS REFERENCEGROUP ONESTIMATES, MODELLING AND PROJECTIONS (2002). Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: Recommendations of the UNAIDS reference group on estimates, modelling and projections. *AIDS* **16** W1–W14.

[46] Alkema L., Raftery A.E., Clark S.J. (2007) Probabilistic Projections of HIV Prevalence Using Bayesian Melding. *The Annals of Applied Statistics* **1** 229–248.

[47] Suh, KN; Kain, KC; Keystone, JS (2004) Malaria *CMAJ* **170** 1693–1702.

[48] Tanner M., Greenwood B., Whitty C.J.M., Ansah E.K., Price R.N., Dondorp A.M., von Seidlein L., Baird J.K., Beeson J.G., Fowkes F.J.I., Hemingway J., Marsh K., Osier F. (2015) Malaria eradication and elimination: views on how to translate a vision into reality. *BMC Med.* **13** 167.

[49] Smith T.A., Chitnis N., Penny M., Tanner M. (2017) Malaria Modeling in the Era of Eradication *Cold Spring Harb Perspect Med.* **7** a025460.

[50] Smith T., Killeen G.F., Maire N,. Ross A., Molineaux L., Tediosi F., Hutton G., Utzinger J., Dietz K., Tanner M. (2006) Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of Plasmodium falciparum malaria: Overview. *Am J Trop Med Hyg* **75** 1–10.

[51] Maire N., Shillcutt S.D., Walker D.G., Tediosi F., Smith T.A. (2011) Cost-effectiveness of the introduction of a pre-erythrocytic malaria vaccine into the expanded program on immunization in sub-Saharan Africa: analysis of uncertainties using a stochastic individual-based simulation model of Plasmodium falciparum malaria. *Value Health.* **15** 1028–38.

[52] Pfeffer D.A., Lucas T.C.D., May D., Harris J., Rozier J., Twohig K.A., Dalrymple U., Guerra C.A., Moyes C.L., Thorn M., Nguyen M., Bhatt S.,Cameron E., Weiss D.J., Howes R.E., Battle K.E.,Gibson H.S., Gething P.W. (2018) malariaAtlas: an R interface to global malariometric data hosted by the

medRxiv preprint doi: https://doi.org/10.1101/2020.08.03.20146241; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Malaria Atlas Project *Malar J.* **17** 352.

[53] Lechthaler F., Matthys B., Lechthaler-Felber G., Likwela J.L., Mavoko H.M., Rika J.M., Mutombo M.M., Ruckstuhl L., Barczyk J., Shargie E., Prytherch H., Lengeler C. (2019) Trends in reported malaria cases and the effects of malaria control in the Democratic Republic of the Congo. *PLoS One* **14** e0219853.

[54] Poole D., Raftery A.E. (2000) Inference for Deterministic Simulation Models: The Bayesian Melding Approach. *Journal of the American Statistical Association* **95** 1244–1255.

[55] Marin J.M., Pudlo P., Sedki M. (2019) Consistency of adaptive importance sampling and recycling schemes. *Bernoulli* **25** 1977–1998.

[56] L. Martino L., Elvira V., Luengo D., Corander J. (2018) Layered adaptive importance sampling. *Statistics and Computing* **27** 599–623.

Department of Plant Sciences
University of Cambridge, Cambridge, UK
rr614@cam.ac.uk

Department of Statistics and
Zeeman Institute
University of Warwick, Coventry, UK
p.touloupou.1@warwick.ac.uk
s.e.f.spencer@warwick.ac.uk

Faculty of Medicine
School of Public Health
Imperial College London, UK
m.basanez@imperial.ac.uk

Big Data Institute
Oxford University, UK
E-mail: deirdre.hollingsworth@bdi.ox.ac.uk
URL: https://www.ntdmodelling.org