

INTEGRATING GEOSTATISTICAL MAPS AND INFECTIOUS DISEASE TRANSMISSION MODELS USING ADAPTIVE MULTIPLE IMPORTANCE SAMPLING

BY RENATA RETKUTE¹, PANAYIOTA TOULOPOU², MARÍA-GLORIA BASÁÑEZ³,
T. DÉIRDRE HOLLINGSWORTH⁴ AND SIMON E. F. SPENCER⁵

¹*Epidemiology and Modelling Group, Department of Plant Sciences, University of Cambridge, rr614@cam.ac.uk*

²*School of Mathematics, University of Birmingham, p.touloupou@bham.ac.uk*

³*London Centre for Neglected Tropical Disease Research and MRC Centre for Global Infectious Disease Analysis, Faculty of Medicine, School of Public Health, Imperial College London, m.basanez@imperial.ac.uk*

⁴*Big Data Institute, Li Ka Shing Centre for Health, Information and Discovery, University of Oxford, deirdre.hollingsworth@bdi.ox.ac.uk*

⁵*Department of Statistics and Zeeman Institute, University of Warwick, s.e.f.spencer@warwick.ac.uk*

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 take the output from a geostatistical model of infectious disease prevalence, incidence or relative risk, and project it forward in time under a mathematical model for transmission dynamics. We adapted the AMIS algorithm so that it can sample from multiple targets simultaneously by changing the focus of the adaptation at each iteration. By comparing our approach against the standard AMIS algorithm, we showed that these novel adaptations greatly improve the efficiency of the sampling. We tested the performance of our algorithm on four case studies: ascariasis in Ethiopia, onchocerciasis in Togo, human immunodeficiency virus (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 (Bhatt et al. (2015)), soil-transmitted helminthiasis (Karagiannis-Voules et al. (2015)), leishmaniasis (Pigott et al. (2014)), onchocerciasis (O’Hanlon et al. (2016)), dengue (Bhatt et al. (2013)), human African trypanosomiasis (Simarro et al. (2012)), HIV (Dwyer-Lindgren et al. (2019)), and diphtheria-pertussis-tetanus vaccine coverage (Mosser et al. (2019)). These maps are made by averaging over many spatially continuous surfaces which are constructed using geopositioned survey data, ecological covariates, and spatial correlations (Amoah, Diggle and Giorgi (2020), Giorgi et al. (2018), Hay et al. (2013)).

Transmission dynamic models have successfully been utilised to evaluate the population dynamics of infectious diseases and assess the impact of interventions (Hollingsworth (2018)). 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.

In this work we develop a statistical framework based on the Adaptive Multiple Importance Sampling algorithm (AMIS; Cornuet et al. (2012)) for effective integration of geostatistical maps and infection transmission models. We investigate the performance of AMIS on

Received November 2020; revised May 2021.

Key words and phrases. Epidemiology, Disease mapping, Parameter estimation, Importance sampling.

four case studies: the soil-transmitted nematode *Ascaris lumbricoides* (causative of ascariasis, a soil-transmitted helminthiasis) in Ethiopia; the filarial parasite *Onchocerca volvulus* (causative of onchocerciasis/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 postcontrol data (i.e., data collected once disease control interventions have been implemented).

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 (AMIS) algorithm (Cornuet et al. (2012)) to target important and underexplored areas of the parameter space based on high-resolution nationwide maps of infection prevalences.

2.1. Disease mapping. Model-based geostatistics allows the study of geographical variation in disease prevalence or incidence, even when the available data are limited to observations from a finite number of sampled locations. High-resolution maps can be constructed using spatially referenced data, including the number of individuals tested, the number who tested positive for disease, and geographical coordinates of the samples (Diggle and Giorgi (2019)). By modelling the spatial-autocorrelation between surveys it becomes possible to interpolate the prevalence between sampled locations. Bayesian methods are frequently employed to fit geostatistical models and provide a posterior distribution over the prevalence map. This is usually represented by M sampled maps capturing the autocorrelation between locations. Further, we will call this set of M maps as our “data.”

2.2. Transmission modelling. Mathematical models of disease transmission are frequently employed to describe and understand infection dynamics, provide predictions of the future burden of disease, and to assess and compare the effectiveness of intervention strategies (Anderson and Basáñez (2015), Anderson and May (1992), Basáñez and Anderson (2016), Keeling and Rohani (2008)). To capture realistic disease dynamics, transmission models may need to describe a range of biological factors and population structures, such as community and age structure, population heterogeneity in risk, vector dynamics, pathogen lifecycles, seasonality, changes immunity or behaviour, and public health interventions. This can lead to very complex models with large numbers of parameters. Fortunately, many of these parameters can be informed from existing scientific literature, laboratory studies, or other data sources, for example, census data. The small number of model parameters remaining can then be fitted to the low-dimensional data on observed cases that are typically available. In this study our focus is on estimating a small number of transmission model parameters that are likely to vary spatially, such as the infection rate, vector-to-host ratio, and heterogeneity or parasite aggregation parameters.

2.3. 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 , that is, $\theta_j \sim \Phi$. The corresponding importance weights are $w_j = \frac{\psi(\theta_j)}{\phi(\theta_j)}$, where $\psi(\theta)$ and $\phi(\theta)$ are, respectively, the target and the proposal density functions (Ripley (1987)).

Importance sampling has been previously used to obtain probabilistic projections of HIV prevalence (Alkema, Raftery and Clark (2007)) and to link geostatistical maps and transmission models of lymphatic filariasis (Touloupou et al. (2020)). Veach and Guibas (Veach and Guibas (1995)) proposed the *deterministic multiple mixture* to pool together importance samples from different proposal distributions. In this case the importance samples $\theta_j^i \sim \Phi_i$, ($1 \leq i \leq I$, $1 \leq j \leq N_i$) and the corresponding importance weights are calculated based on the mixture of weights (Cornuet et al. (2012)), given by

$$(2.1) \quad w_j^i = \frac{\psi(\theta_j^i)}{\frac{1}{N_1 + \dots + N_I} \sum_{l=1}^I N_l \phi_l(\theta_j^i)}.$$

This idea was extended by Cornuet and coauthors (Cornuet et al. (2012)) and Raftery and Bao (2010), who proposed using the deterministic multiple mixture formula to construct importance proposals, sequentially and adaptively. First, the importance weights for current iteration, i , are calculated, while the importance weights for previous iterations, $1 \leq u \leq i - 1$, are recalculated, based on all proposals up to the iteration i . Second, samples from all iterations are used to construct the next proposal distribution.

Adaptive multiple importance sampling has been utilised in a variety of research fields, including population genetics (Siren, Martinen and Corander (2010)), environment illumination computations (Sbert and Havran (2017)), and signal communications (Elvira and Santamaria (2019)).

The novelty of our study comprises: (i) applying the AMIS approach to real-world geospatial data; (ii) applying AMIS to multiple targets simultaneously 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 a theoretical and a practical perspective.

2.4. 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, for example, infection prevalences, annual incidence, or number of cases. From here onward, we will assume that we are dealing with the prevalence of infection. Suppose we have an observed prevalence matrix $Q = (q_{l,m})_{L \times M}$, 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 (Anderson and Basáñez (2015), Anderson and May (1992), Basáñez and Anderson (2016), Keeling and Rohani (2008)). Suppose that 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 , that is, $F(\theta_j) = p_j$.

At iteration i , N_i parameters are sampled from a proposal density ϕ_i and denoted by θ_j , for $j = \sum_{u=1}^{i-1} N_u + 1, \dots, \sum_{u=1}^i N_u$. In the first iteration we use the prior as the proposal, so $\phi_1(\theta) = \pi(\theta)$. Second, the transmission model is used to calculate the prevalence corresponding to the sampled parameters, given by $p_j = F(\theta_j)$. Next, parameter vector j is weighted by its importance weight and then reweighted for every pixel l so that the weighted distribution of simulated prevalences resembles the distribution of observed prevalences at that pixel. This reweighting is performed using an empirical version of the usual change of measure formula (Radon–Nikodym derivative), as described in Touloupou et al. (2020). The weight for parameter vector j in pixel l at iteration i is, therefore, given by

$$w_{lj}^i = \frac{f_l(p_j)}{g(p_j)} \frac{\pi(\theta_j)}{\frac{1}{N_1 + \dots + N_i} \sum_{u=1}^i N_u \phi_u(\theta_j)},$$

where $f_l(p_j)$ 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 (Touloupou et al. (2020)),

$$f_l(p_j) = \frac{1}{\delta M} \sum_{m=1}^M \mathcal{I}(p_j - \delta/2 \leq q_{l,m} \leq p_j + \delta/2),$$

$$g(p_j) = \frac{1}{\delta K} \sum_{u=1}^{N_1+\dots+N_i} \frac{\mathcal{I}(p_j - \delta/2 \leq p_u \leq p_j + \delta/2)\pi(\theta_u)}{\frac{1}{N_1+\dots+N_i} \sum_{v=1}^i N_v \phi_v(\theta_u)},$$

where $K = \sum_{u=1}^{N_1+\dots+N_i} \frac{\pi(\theta_u)}{\frac{1}{N_1+\dots+N_i} \sum_{v=1}^i N_v \phi_v(\theta_u)}$.

These weights are normalised to give $\hat{w}_{lj}^i = \frac{w_{lj}^i}{\sum_{u=1}^{N_1+\dots+N_i} w_{lu}^i}$ for $l \in \{1, \dots, L\}$ and $j \in \{1, \dots, N_1 + \dots + N_i\}$.

We use Kish’s effective sample size (ESS; Kish (1965)) as a measure of the quality of representation of the pixel’s prevalence distribution by the simulations. We denote the ESS for pixel l after iteration i as

$$ESS_l^i = \left(\sum_{j=1}^{N_1+\dots+N_i} (\hat{w}_{lj}^i)^2 \right)^{-1}.$$

We set a required ESS for all of the pixels, denoted by ESS^R . We call pixels “active” if they have an ESS below the target and use the weights for the active pixels to design the proposal for the next iteration of the AMIS algorithm. This targets sampling toward areas of the parameter space that benefit pixels that have not yet reached the required ESS. Let the index set of the active pixels be denoted by $A_i = \{l : ESS_l^i < ESS^R\}$.

More precisely, we use the mean weight of each simulation over the active set to determine the next proposal distribution. Let $\bar{w}_j^i = \frac{1}{|A_i|} \sum_{l \in A_i} \hat{w}_{lj}^i$ be the mean weight in the active set for simulation j . A suitable proposal can be found by fitting a density to the weight samples (θ_j, \bar{w}_j^i) , for $j \in \{1, \dots, N_1 + \dots + N_i\}$. The algorithm continues until all pixels meet the ESS requirement or the maximum number of iterations, I , have been completed. Pseudo code for the algorithm is shown in Algorithm 1.

2.5. Notes on Algorithm 1.

1. In practice, the normalising constants involving δ , M , and K on the densities $f_l(p_j)$ and $g(p_j)$ in the empirical Radon–Nikodym derivative do not need to be calculated, as they disappear when the weights are normalised in Step 5.

2. We have added a maximum number of iterations to the AMIS algorithm because otherwise it is not guaranteed to finish, at least in theory. As new simulations are added, it is possible for the ESS in a particular pixel to go down, and so pixels may re-enter the active set, even after they have previously met the required ESS. This happens when Kish’s ESS formula provides a poor estimate, for example, when all of the simulations are given equally low weights. When new simulations arrive in a high weight region, the ESS estimate drops sharply, but these new simulations provide guidance for future proposals. In practice, we have not observed any problems in getting all of the pixels to meet the ESS requirement.

3. For the algorithm to start successfully, it is desirable for the first iteration to provide simulations covering the full support of the prevalence distributions from the geostatistical mapping. Otherwise, there are regions of the prevalence distribution that do not have

Algorithm 1 AMIS for integrating geostatistical maps and transmission models

- 1: Set $i = 1$, and let $A_1 = \{1, \dots, L\}$ be the set of active pixels.
- 2: Sample N_i parameter vectors from proposal density $\phi_i(\theta)$ and label them θ_j , for $j = \sum_{u=1}^{i-1} N_u + 1, \dots, \sum_{u=1}^i N_u$. If $i = 1$, ϕ_i is the prior distribution: $\phi_1(\theta) = \pi(\theta)$. If $i > 1$, ϕ_i is a density fitted to the weighted sample from the previous iteration: $\{(\theta_j, \bar{w}_j^{i-1}) : j = 1, \dots, (N_1 + \dots + N_{i-1})\}$.
- 3: Simulate the transmission model for the new parameter vectors: $p_j = F(\theta_j)$.
- 4: Calculate weights for parameter vectors $j \in \{1, \dots, N_1 + \dots + N_i\}$, for each pixel with index $l \in \{1, \dots, L\}$,

$$w_{lj}^i = \frac{f_l(p_j)}{g(p_j)} \frac{\pi(\theta_j)}{\frac{1}{N_1 + \dots + N_i} \sum_{u=1}^i N_u \phi_u(\theta_j)},$$

where

$$f_l(p_j) = \frac{1}{\delta M} \sum_{m=1}^M \mathcal{I}(p_j - \delta/2 \leq q_{l,m} \leq p_j + \delta/2),$$

$$g(p_j) = \frac{1}{\delta K} \sum_{u=1}^{N_1 + \dots + N_i} \frac{\mathcal{I}(p_j - \delta/2 \leq p_u \leq p_j + \delta/2) \pi(\theta_u)}{\frac{1}{N_1 + \dots + N_i} \sum_{v=1}^i N_v \phi_v(\theta_u)}.$$

Here, \mathcal{I} is the indicator function and $K = \sum_{u=1}^{N_1 + \dots + N_i} \frac{\pi(\theta_u)}{\frac{1}{N_1 + \dots + N_i} \sum_{v=1}^i N_v \phi_v(\theta_u)}$.

- 5: Normalise weights for each pixel $l \in \{1, \dots, L\}$,

$$\hat{w}_{lj}^i = \frac{w_{lj}^i}{\sum_{u=1}^{N_1 + \dots + N_i} w_{lu}^i}.$$

- 6: Calculate effective sample size (ESS) for each pixel $l \in \{1, \dots, L\}$,

$$ESS_l^i = \left(\sum_{j=1}^{N_1 + \dots + N_i} (\hat{w}_{lj}^i)^2 \right)^{-1}.$$

- 7: Update the set of active pixels, $A_i = \{l : ESS_l^i < ESS^R\}$.
- 8: Calculate the average weight of parameter vectors based on individual weights of the active pixels,

$$\bar{w}_j^i = \frac{1}{|A_i|} \sum_{l \in A_i} \hat{w}_{lj}^i.$$

- 9: Set $i = i + 1$.
 - 10: Repeat from step 2 until A_i is empty or I iterations have been completed.
-

simulations to represent them. This limitation means that our approach will not succeed for high-dimensional parameter spaces.

4. Any attempt to represent a probability distribution by weighted samples is necessarily going to be an approximation. There are many possible choices for the form of the empirical Radon–Nikodym derivative in Step 4 of the algorithm, and several options are discussed in Touloupou et al. (2020), including some variants that minimise certain measures of disagreement and others that are faster to compute.

5. In most settings the largest computational cost associated with the algorithm will be in simulating from the transmission model for each parameter vector. Our algorithm is designed

to reduce these costs by targeting simulations toward areas of the parameter space where they will be most useful.

6. Although our algorithm is described in terms of a deterministic transmission model $F(\theta)$, it can also be applied to stochastic models by supplementing each parameter vector with an element ω from the probability space describing the stochasticity in the transmission model. This leads to a deterministic relationship $p = F(\theta, \omega)$. We illustrate this in the Supplementary Material (Retkute et al. (2021)), Section C, with a stochastic transmission model for malaria.

2.6. General implementation of Algorithm 1. To generate the adaptive proposals in the examples that follow, we fitted a mixture of Gaussian distributions to the weighted samples. We then set the next proposal density to be the equivalent mixture of t -distributions with three degrees of freedom. 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 (Cappé et al. (2008)). We used the R package `mclust` to fit the proposal distribution (Scrucca et al. (2016)). The package uses an optimal number of cluster components based on Bayesian information criterion (BIC; Fraley and Raftery (2003)).

2.7. Incorporating data from multiple time points. Our proposed framework can be naturally extended to the case where geostatistical maps are available at multiple time points $t = 1, \dots, T$. If we assume that the geo-statistical maps from each time point are produced independently, weights for pixel $l \in \{1, \dots, L\}$ and parameter vector $j \in \{1, \dots, N_1 + \dots + N_i\}$ will be equal to

$$w_{lj}^i = \left(\prod_{t=1}^T \frac{f_{l,t}(p_{j,t})}{g_t(p_{j,t})} \right) \frac{\pi(\theta_j)}{\frac{1}{N_1 + \dots + N_i} \sum_{u=1}^i N_u \phi_u(\theta_j)}.$$

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.

Furthermore, if the geostatistical maps are not independent but are fitted to multiple time points jointly, the original weighting formula can be used to capture the spatial dependence between maps, simply by taking p_j to be the vector of prevalences from simulation j across time. However, producing transmission model simulations that match multiple time points will be extremely challenging and is likely to be an area for future research.

3. Applications. We applied AMIS to four case studies: ascariasis in Ethiopia, onchocerciasis in Togo, HIV infection in Botswana, and malaria in the DRC. The framework of the AMIS approach, and the description and objectives of each case study are shown in Figure 1. To set up a case study, three components are required in order to run the AMIS framework: geostatistical map, parameter prior distribution, and transmission model. We used survey data (Pullan et al. (2014)) to produce a series of prevalence maps of *Ascaris lumbricoides* infection with different spatial resolutions and numbers of samples of the observed prevalences at each pixel. For onchocerciasis, HIV and malaria, we used publicly available maps (Bhatt et al. (2015), Dwyer-Lindgren et al. (2019), O’Hanlon et al. (2016)). 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 coverage of artemisinin combination therapies (ACTs) was very low in the DRC (Bhatt et al. (2015), Ntamabyaliro et al. (2018)). The analyses for HIV and malaria are presented in the Supplementary Material, Sections B and C.

For all four case studies we set $\delta = 0.05$, $N_i = 1000$, $I = 30$, and $\text{ESS}^R = 2000$. Therefore, the maximum number of sampled parameter vectors was 30,000.

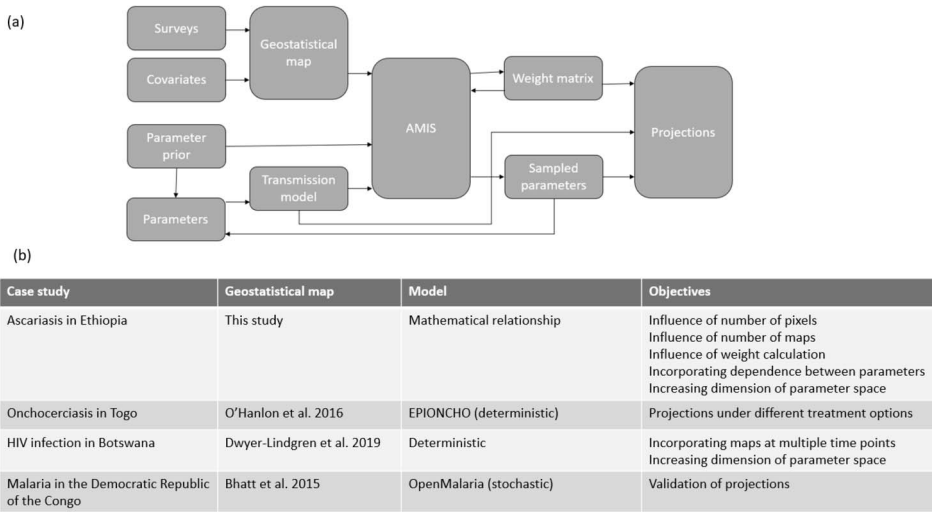


FIG. 1. (a) AMIS framework. (b) Description of case studies. The results of the ascariasis and onchocerciasis analyses are presented in the main text; those for HIV and malaria are in the Supplementary Material.

3.1. *Ascariasis in Ethiopia.* *Ascaris lumbricoides* is an intestinal nematode, also known as roundworm (Norhayati et al. (2003)). It is estimated that 819 million people worldwide are infected by *A. lumbricoides* (Pullan et al. (2014)).

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 (Anderson and May (1992))

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

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 (Anderson and May (1985)).

Data on *Ascaris* surveys for Ethiopia were downloaded from the Global Atlas of Helminth Infection (London Applied & Spatial Epidemiology Research Group, LASER; Pullan et al. (2014)). The data consisted of longitude, latitude, number of school-aged children (SAC; 5 to 14-years-old) examined, number of positives (for adult worm presence and prevalence prior to wide-spread deworming programmees (Pullan et al. (2014)). Entries with missing geolocation or prevalence values were excluded from further analysis, leaving 290 survey values. We used the INLA-R package (Rue, Martino and Chopin (2009)) and a stochastic partial differential equation approach (Lindgren, Rue and Lindström (2011)) to produce national level prevalence maps. We incorporated spatial correlation (nonstationary locally isotropic stochastic partial differential equation/Gaussian Markov random field model) and included elevation as an environmental covariate (Brooker, Clements and Bundy (2006)) downloaded from the raster package (Hijmans (2019)).

We mapped *Ascaris* prevalence in Ethiopia at resolutions 5 km × 5 km ($L = 37,695$ pixels) and 10 km × 10 km ($L = 11,369$ pixels) and sampled 100, 500, 1000 and 2000 individual maps from the posterior distribution for both resolutions. A map showing communities sampled in Ethiopia (Pullan et al. (2014)) and triangulated mesh used to build a geostatistical

model of *Ascaris* prevalence in Ethiopia (Rue, Martino and Chopin (2009)) are shown in Supplementary Material Figure S1.

3.1.2. Implementation of AMIS. A consistent relationship between mean worm burden and prevalence of *Ascaris lumbricoides* infection has been observed in a dataset collated from a range of geographically distinct human communities (Guyatt et al. (1990)). Therefore, we have tested 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 (Guyatt et al. (1990)). This has led to the following joint prior for the parameters: 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 $k|W \sim N(0.3337 + 0.0171 \times W, \sigma(W)^2)$ (Supplementary Material Figure S2). Although typical values for mean worm burden are 10–20 (Fowler and Hollingsworth (2016)), much higher numbers have been observed in field conditions (Anderson and May (1992), Guyatt et al. (1990)). For details on fitting procedure, see Appendix A in the online Supplementary Material. The estimated relationship between mean worm burden and degree of parasite aggregation is shown in Supplementary Material Figure S2.

3.1.3. Results. Details of the map and the parameters sampled for resolution $5 \text{ km} \times 5 \text{ km}$ and $M = 2000$ are shown in Figure 2. First, we plotted a map of mean prevalence for each pixel (Figure 2(a)) which ranged between 0.1 and 0.55. Second, a histogram of sampled prevalences, colour coded according to a proportion sampled at each iteration, is shown in Supplementary Material Figure S3(a). Here, iterations 2 and above targeted prevalences, mostly, between 0 and 0.5, as this is the range of mapped prevalences of *Ascaris* (Supplementary Material Figure S1(b)). Third, Figure 2(b) shows prevalence as a function of log mean number of worms, $\log(W)$, and degree of parasite aggregation, k ; this function has a complex 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, that is, along this contour (Figure 2(c)).

Supplementary Material Figure S4 shows how sampled values of $\log(W)$ change with each iteration. Prevalences corresponding to sampled parameters have good agreement with observed values from Guyatt et al. (1990).

We tested two alternative methods: based on sampling from the prior only and based on AMIS with $|A_i| = L$ for all iterations. 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.,

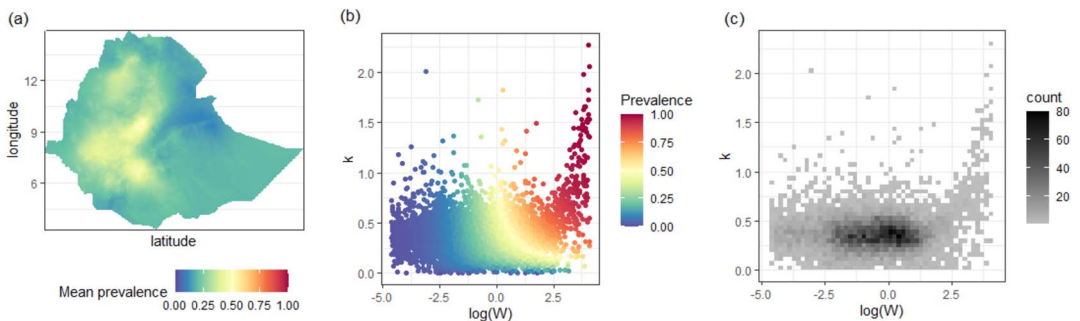


FIG. 2. Results for *Ascaris lumbricoides* in Ethiopia. (a) Map of mean prevalences constructed from survey data in Ethiopia (Pullan et al. (2014)). The resolution is $5 \text{ km} \times 5 \text{ km}$ and $M = 2000$. Sampled communities and prevalences are shown in Figure S1. (b) Scatter plot of sampled model parameters and corresponding prevalence. (c) Density of sampled parameters. Prevalence and density distributions are shown as a function of log mean number of worms, $\log(W)$, and degree of parasite aggregation, k .

TABLE 1

Number of iterations to achieve $\min(\text{ESS}) \geq \text{ESS}^R$, with $\text{ESS}^R = 2000$. Here, P stands for sampling from the prior only, S stands for standard AMIS with $|A_i| = L$ for all iterations, and A is based on Algorithm 1. Each iteration samples 1000 parameter sets

L	$M = 100$			$M = 500$			$M = 1000$			$M = 2000$		
	P	S	A	P	S	A	P	S	A	P	S	A
11,369	19	26	9	19	25	9	18	26	9	17	26	8
37,695	20	26	9	19	25	9	18	26	9	18	27	9

with set of active pixels updated every iteration) outperformed the other two methods and required around a third of sampled parameters. The two alternative methods produced samples from different regions of parameter space (Supplementary Material Figure S5). When sampling from the prior, we found that a histogram of posterior prevalences had a “U-shape,” that is, very low and very high values of prevalences are oversampled which does not correspond to the mapped prevalences (Supplementary Material, Figure S1(b)). When sampling with the set of active pixels equal to all pixels, the algorithm targeted parameter regions similar to those of Algorithm 1 but still required twice as many iterations.

3.2. *Extending the algorithm for higher dimensions.* We used a modified *A. lumbricoides* model, defined in Section 3.1, to investigate how the AMIS algorithm performs when the number of parameters increases. We assumed that the parasite has $d - 1$ developmental stages or different phenotypes, and that this has an additive effect on disease prevalence, that is, the mean number of worms per host is equal to a sum of the mean number of worms in each developmental stage per host,

$$(3.2) \quad P = 1 - \left(1 + \frac{\sum_{i=1}^{d-1} W_i}{k} \right)^{-k}.$$

This gives a total of d parameters (i.e., the mean number of worms for each group $d - 1$ and the degree of clumping of parasites within a population of hosts, k). To keep this illustrative example as close to a real application as possible, we have run the analysis on a map of *Ascaris* prevalence in Ethiopia (10 km \times 10 km resolution, $L = 11,369$ pixels). We used the following joint prior for the parameters: $\log(W_i) \sim U[\log(0.01), \log(60)]$, and a Gaussian distribution for the degree of clumping,

$$k|(W_1, \dots, W_{d-1}) \sim N\left(0.3337 + 0.0171 \times \sum_{i=1}^{d-1} W_i, \sigma\left(\sum_{i=1}^{d-1} W_i\right)^2\right).$$

We have run the AMIS algorithm with two fixed sample sizes: $N_i = 1000$ and $N_i = 10,000$ and a sample size which increases with dimension $N_i = d \times 1000$. Table 2 shows how the number of iterations and total computational time (in hours) changed when the number of parameters d was increased from two to 10. Calculations were run on a single Macbook Pro with 3.5 GHz six-core Intel Processor. For this particular model, running simulations is computationally cheap—to run 1000 simulations, it takes approximately two seconds. However, each iteration requires calculating weights for $L = 11,369$ pixels across all sampled parameter sets. The number of iterations required to achieve $\text{ESS} \geq 2000$ increased from nine to 75 for $N_i = 1000$, while run times increased from roughly half an hour to 52 hours. When we increased the sample size to 10,000 parameter sets per iteration, the number of iterations and computational time decreased considerably. Increasing the number of simulations with increasing parameter dimension led to an increasing number of iterations with a greater than

TABLE 2
 Results the modified *A. lumbricoides* model with d parameters: Number of iterations required to achieve ESS > 2000 and total computational time

N_i	No. param.	2	3	4	5	6	7	8	9	10
1000	Iterations	9	11	14	15	36	41	52	63	75
	Simulations	9000	11,000	14,000	15,000	36,000	41,000	52,000	63,000	75,000
	Time (h)	0.65	0.99	1.60	1.78	10.97	14.44	23.4	35.6	52.4
10,000	Iterations	2	2	3	5	6	6	7	7	9
	Simulations	20,000	20,000	30,000	50,000	60,000	60,000	70,000	70,000	90,000
	Time (h)	0.46	0.47	0.94	2.44	3.46	3.47	4.61	4.61	6.01
$d \times 1000$	Iterations	5	4	5	8	8	7	9	9	9
	Simulations	10,000	12,000	20,000	40,000	48,000	49,000	72,000	81,000	90,000
	Time (h)	0.42	0.42	0.83	2.57	3.19	2.95	5.54	6.34	6.01

linear increase with dimension. For example, when running the algorithm for the model with 10 parameters, the number of iterations was reduced from 75 to nine, while the total timing was reduced from 52 hours to six hours.

This case study indicates that the algorithm can be run successfully in higher dimensions, but doing so requires more than linear increases in the number of simulations. In this example the model simulations were very fast to compute, and so most of the computational burden came from weighting the simulations and calculating the ESS. For more complex and time-consuming simulation models, a smaller N_i will be needed to minimise computational time.

3.3. 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 (Colebunders et al. (2018)). Programs for the control and elimination of onchocerciasis have been targeted to the areas most affected and great strides have been made, but challenges remain to achieve elimination of transmission (Colebunders et al. (2018)). 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; Plaisier et al. (1991)). Annual or biannual 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 (Hougard et al. (2001)). 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 (Cheke et al. (2015), Jacob et al. (2018), Routledge et al. (2018)).

3.3.1. Model and data. We simulated onchocerciasis transmission in communities using the deterministic version of the EPIONCHO model (Walker et al. (2017)). 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 nonfertile 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 worms, were set as in the supplementary material of Walker et al. (2017). 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 5 km \times 5 km spatial resolution developed in O'Hanlon et al. (2016) for

the area of the Onchocerciasis Control Programme in West Africa (OCP). The data for Togo consist of $M = 2000$ samples for each pixel, with each individual prevalence map consisting of $I = 9360$ pixels.

3.3.2. 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 worms $k \sim U[0, 3]$.

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 during the OCP (1974-2002) and in Special Intervention Zones (after OCP's closure in 2002), we have focused on ivermectin MDA because no large-scale vector control is currently implemented. As ivermectin is mostly microfilaricidal, and the microfilariae (the adult parasite larval progeny) 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 (Walker et al. (2017)). 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 alternative treatment strategy (Verver et al. (2018))), the latter meaning that treatment is reaching nearly the entirety of the eligible population. In both cases the proportion of systematic nonadherers was set at 5%.

3.3.3. Results. Similar to the case of ascariasis (Section 3.1), there is a strong heterogeneity in the spatial distribution of onchocerciasis prevalence in Togo (Figure 3(a)). The mean microfilarial prevalence can reach levels of up to 80% in holoendemic areas with high vector biting rates. Figures 3(b)–(c) show sampled parameter vectors and simulated prevalences. The AMIS framework required 10 iterations to achieve $ESS \geq ESS^R$ in all pixels (Supplementary Material, Figure S3(b)). 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 (Supplementary Figure S3(b)).

We compared the impact of 15 years of annual ivermectin MDA with total population coverage 65% and 80%. Figure 4 shows the reduction in microfilarial prevalence in the endemic areas five, 10, and 15 years after the start of MDA. Pre-control endemicity levels (i.e., preintervention infection prevalence and intensity) have been indicated as a crucial factor determining the success of intervention strategies to achieve elimination of transmission

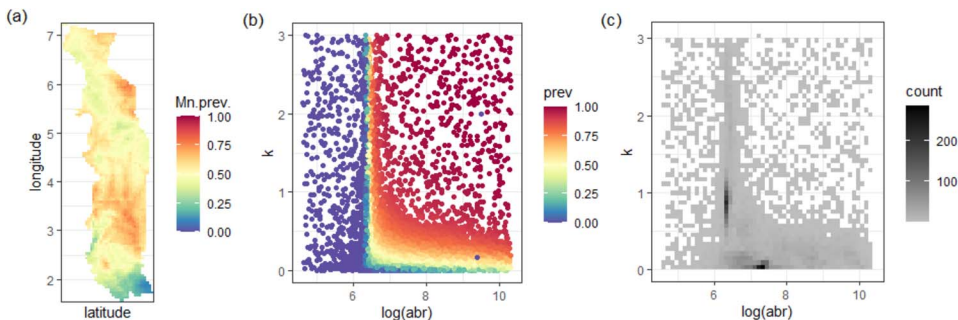


FIG. 3. Results for onchocerciasis in Togo. (a) Map of mean baseline prevalence (O’Hanlon et al. (2016)). (b) Scatter plot of sampled model parameters and corresponding prevalence. (c) Density of sampled parameters. In (b) and (c), prevalence and density distributions are shown as a function of log annual biting rate, $\log(abr)$, and the aggregation of adult worms, k .

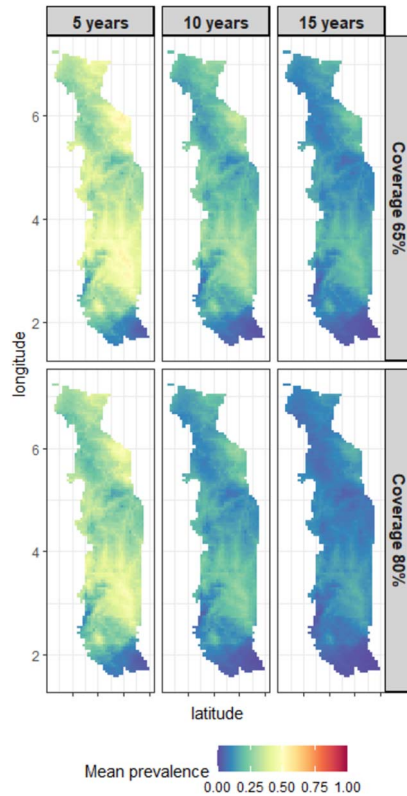


FIG. 4. Comparison of the impact of ivermectin coverage levels on elimination of onchocerciasis in Togo. The interventions compared are mass drug administration of ivermectin at 65% coverage and mass drug administration at 80% coverage. Projections were run using the EPIONCHO model (Walker et al. (2017)).

(Verver et al. (2018)), 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 (i.e., the ability to bounce-back following treatment) was markedly higher for stronger levels of exposure heterogeneity, that is, lower values of k (Hamley et al. (2019)). Figure 5 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 external data to restrict parameter priors (as has been done in Section 3.1) or using postcontrol maps within the AMIS algorithm (as was done in the HIV example described in Section C of the Supplementary Material).

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 infection transmission models. We have selected case studies that span from viruses (HIV), protozoan parasites (malaria), to metazoan parasites with direct (ascariasis)

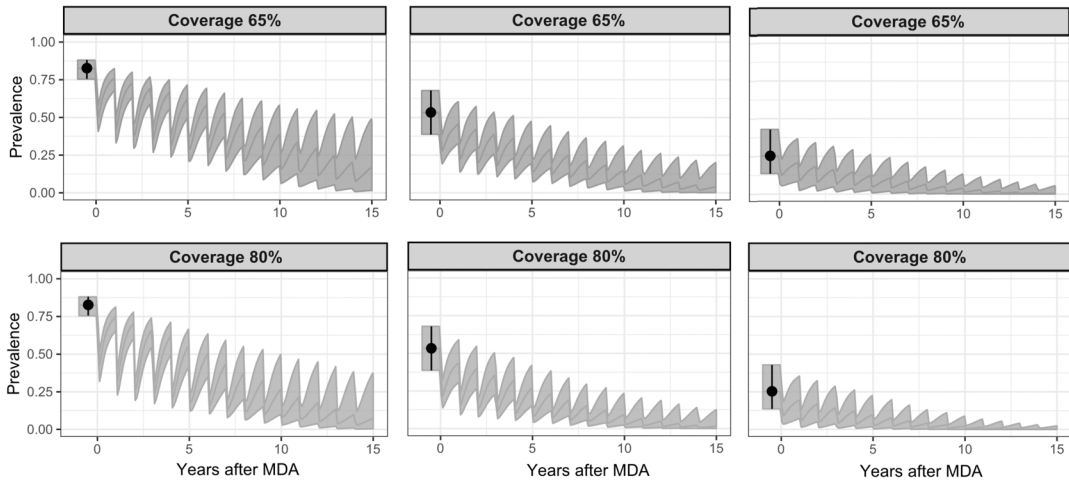


FIG. 5. Comparison of two strategies: Minimum recommended 65% coverage of total population (top row) and enhanced 80% coverage as an alternative treatment strategy (bottom row) for onchocerciasis control and elimination at three levels of baseline endemicity: hyperendemicity (left column), mesoendemicity (middle column), and hypoendemicity (right column). Projections were simulated using the EPIONCHO model (Walker *et al.* (2017)). Mapped modelled prevalences are shown in black (O'Hanlon *et al.* (2016)). Figures show mean and 95% interquartile range.

and indirect, vector-borne, life-cycles (onchocerciasis) to illustrate the applicability of the approach in major infectious diseases of humans. Below we discuss the benefits and limitations of the methodology.

4.1. *Parameter estimation based on prevalences at a national level.* The diversity of infectious diseases studied and the range of mathematical models used, allowed us to explore the performance of our framework under different conditions. As the results 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 in which 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 very 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 all pixels in the active set ($|A| = L$) for all iterations. As can be seen from Table 1, the method based on weights of active pixels required two to three times fewer iterations than 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 toward the pixels that have already been well served by earlier iterations of the algorithm and not toward the required posterior prevalences for all of the pixels, for example simulated prevalences between 0.5 and 0.75 (Supplementary Material, Figure S5).

4.3. *Influence of size of geospatial maps.* In Section 3.1 we found that the number of pixels (i.e., $L = 11,369$ pixels and $L = 37,695$ pixels for the prevalence of *Ascaris* in school-aged children in Ethiopia) had little influence on the performance of AMIS 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 eight or nine when the number of sampled 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 sampled maps will not require more model simulations but, in fact, can decrease this number in comparison to a smaller number of maps. Having a smaller number of sampled 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 infeasible 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, where parameters for the degree of parasite aggregation were sampled from the Gaussian distribution conditioned on the values of mean number of worms.

4.5. Impact of intervention programs. In Section 3.3 we have sampled parameters for onchocerciasis in Togo, simulated the transmission model further forward in time under two levels of therapeutic 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 via the simulations and produced a wide range of future projected prevalences. Uncertainty can be reduced by utilizing postcontrol maps within the AMIS algorithm.

4.6. Comparison with importance sampling. Previously, importance sampling was applied to investigate the impact of intervention programs for lymphatic filariasis in seven countries in Africa (Touloupou et al. (2020)). The model for lymphatic filariasis transmission had four parameters; 100,000 model simulations were generated with parameters drawn from prior distributions. The prevalence map had 114,667 pixels in the study area. The importance sampling method led to ESS values in a range between 1 and 3500, that is, a fraction of pixels had very low ESS values. Our work extends the methodology introduced in Touloupou et al. (2020), in a sense that both studies use the same form of the empirical Radon–Nikodym derivative, that is, how to reweight the simulations to match pixel prevalence distributions. However, the aim of this study is to apply a version of the AMIS algorithm to greatly reduce the overall number of simulations required while achieving sufficient ESS values for all pixels.

4.7. Comparison with Bayesian melding. A probabilistic approach, called Bayesian melding, combines via logarithmic pooling two priors, one implicit and one explicit, on each output (Poole and Raftery (2000)). Bayesian melding has been proposed and used to account for uncertainty in parameters and model projections for HIV (Alkema, Raftery and Clark (2007)) and filarial infections (Michael et al. (2018)). In the case of the HIV study, 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 (Alkema, Raftery and Clark (2007)). These data would be equivalent to a single pixel in our case. As shown in Section B in the Supplementary Material, the AMIS algorithm offers improved computational efficiency and higher resolution in comparison to this application of Bayesian melding. In the case of filarial infections, the initial stage of the algorithm used prior distributions to generate 200,000 parameter vectors, and 500 parameter sets were sampled with probabilities proportional to their relative log-likelihood values for each of 15 endemic sites in Uganda and Nigeria (Michael et al. (2018)). Our results for onchocerciasis in Togo, presented in Section 3.3, showed that 11,000 parameters sets were enough to guarantee ESS values above 2000 for the map consisting of 29,274 pixels.

4.8. *Using postcontrol maps.* For the malaria case study in Section C of the Supplementary Material, our results demonstrate that, when sampling parameters using baseline as well as postcontrol maps, care has to be taken that the simulated epidemiological curves are able to support the observed dynamics. We assumed that 15% coverage of insecticide treated nets (ITNs) was achieved by 2010 (Table S1), but this coverage would not reduce mean prevalence from 0.75 to 0.25, as seen in Figure S10. The AMIS framework produced a trade-off between fitting to maps in 2005 and 2010. However, the schedule of ITN 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.

4.9. *Time requirements for the algorithm.* The computational cost of the AMIS algorithm has the following components: (i) performing model simulations for a parameter vector; (ii) sampling parameters from the prior or the mixture, and (iii) calculating the weights for each pixel across all parameter sets, and the corresponding ESS.

We have investigated models with wide range of timings for individual runs: to produce 1000 simulations it takes approximately two seconds for the ascariasis model, 35 minutes for the HIV model, 4.6 hours for the onchocerciasis model, and up to 12 hours for the malaria model.

In Section 3.2 we found that computational time increased quadratically with every iteration (Table 2). For the ascariasis case study with the fastest computational time for the model simulations, nine AMIS iterations could be done in about half an hour, while 36 iterations would take 11 hours. Therefore, minimising the number of samples and number of iterations will be crucial when simulating from the model is computationally demanding. However, producing the model simulations within each AMIS iteration is easily parallelisable. An alternative version of the empirical Radon–Nikodym derivative, which is faster to compute for large numbers of samples, is discussed in Touloupou et al. (2020).

4.10. *Tuning of algorithm 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 (ESS). Therefore, the performance of the algorithm depends on the choice of threshold ESS^R . In our case studies we have set $ESS^R = 2000$. When calculating the weight is computationally expensive, lower values of ESS^R might be required.

We have sampled from prior distribution only in the first AMIS iteration. In Siren, Marttinen and Corander (2010), the authors set half of the parameters to be sampled from the prior. In our case we do not know in advance how many iterations will be required to have enough parameter sets to satisfy the condition $\min(\text{ESS}) \geq ESS^R$. However, our results for all case studies indicate that sampling 1000 parameter sets was enough to gauge important regions of parameter space for further analysis, at least in two to four dimensions. We also fixed the

number of samples per iteration $N_i = 1000$. We anticipate that decreasing the value of N_i could lead to a lower number of parameter sets required to achieve $\min(\text{ESS}) \geq \text{ESS}^R$, 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, but this would get increasingly difficult as the dimension of the parameter space increases. 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 $\min(\text{ESS}) = 67.4$ and $\max(\text{ESS}) = 435.1$ after the first iteration, which are high in comparison to a number of sampled parameters, that is, 1000.

4.11. Alternative AMIS procedures. Recently, a modified AMIS has been proposed, where an importance sampling distribution at iteration i is built based on samples from iteration $i - 1$ but weights for all iterations are recalculated after the last iteration (Marin, Pudlo and Sedki (2019)). A simpler recycling strategy could offer computational savings, but this would lose the advantage of the adaptive nature of the AMIS framework we proposed, that is, to utilise information on the ESS, based on all sampled parameter vectors rather than 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 (Martino et al. (2017)). Parallelisation might be a promising route, for example, by sampling N_i/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_i/d parameter vectors.

5. Conclusions. Infectious diseases remain an important health problem worldwide. We have introduced a novel method to sample parameters of infectious disease transmission models given high-resolution prevalence maps. Our strategy of using an adaptive mixture of importance proposals based on a transmission model's similarity to mapped prevalence distributions leads to an efficient exploration of the parameter space, at least in low dimensions. 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 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. Source code for replication is available in the Supplementary Material (Retkute et al. (2021)) and online at <https://github.com/rretkute/AMISEpi>.

Acknowledgments. 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.

Funding. The authors gratefully acknowledge funding of the NTD Modeling Consortium by the Bill and Melinda Gates Foundation [OPP1184344, OPP1186851, OPP1156227]. MGB acknowledges joint centre funding [MR/R015600/1] by the U.K. Medical Research Council (MRC) and the U.K. Department for International Development (DFID) under the MRC/DFID Concordat agreement which is also part of the European and Developing Countries Clinical Trials Partnership (EDCTP2) program supported by the European Union.

SUPPLEMENTARY MATERIAL

Supplement to “Integrating geostatistical maps and transmission models using adaptive multiple importance sampling” (DOI: [10.1214/21-AOAS1486SUPPA](https://doi.org/10.1214/21-AOAS1486SUPPA); .pdf). Additional figures for *A.lumbricoides* and the analyses for HIV and malaria are provided in the online supplement.

Source code for “Integrating geostatistical maps and transmission models using adaptive multiple importance sampling” (DOI: [10.1214/21-AOAS1486SUPPB](https://doi.org/10.1214/21-AOAS1486SUPPB); .zip). R source code for the AMIS framework, disease models and prevalence data described in this paper.

REFERENCES

- ALKEMA, L., RAFTERY, A. E. and CLARK, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *Ann. Appl. Stat.* **1** 229–248. MR2393849 <https://doi.org/10.1214/07-AOAS111>
- AMOAH, B., DIGGLE, P. J. and GIORGI, E. (2020). A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. *Biometrics* **76** 158–170. MR4098552 <https://doi.org/10.1111/biom.13142>
- ANDERSON, R. M. and BASÁÑEZ, M. G., eds. (2015). Adv Parasitol., Part A. In *Mathematical Models for Neglected Tropical Diseases: Essential Tools for Control and Elimination* **87** Academic Press, Oxford.
- ANDERSON, R. M. and MAY, R. M. (1985). Helminth infections of humans: Mathematical models, population dynamics, and control. *Adv. Parasitol.* **24** 1–101. [https://doi.org/10.1016/s0065-308x\(08\)60561-8](https://doi.org/10.1016/s0065-308x(08)60561-8)
- ANDERSON, R. A. and MAY, R. M. (1992). *Infectious Diseases of Humans*, 1st ed. ed. Oxford Univ. Press, Oxford.
- BASÁÑEZ, M. G. and ANDERSON, R. M., eds. (2016). Adv Parasitol., Part B. In *Mathematical Models for Neglected Tropical Diseases: Essential Tools for Control and Elimination* **94** Academic Press, Oxford.
- 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. et al. (2013). The global distribution and burden of Dengue. *Nature* **496** 504–507.
- BHATT, S., WEISS, D. J., CAMERON, E., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., BATTLE, K., MOYES, C. L., HENRY, A. et al. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526** 207–211. <https://doi.org/10.1038/nature15535>
- BROOKER, S., CLEMENTS, A. C. A. and BUNDY, D. A. P. (2006). Global epidemiology, ecology and control of soil-transmitted helminth infections. *Adv. Parasitol.* **62** 221–261.
- CAPPÉ, O., DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2008). Adaptive importance sampling in general mixture classes. *Stat. Comput.* **18** 447–459. MR2461888 <https://doi.org/10.1007/s11222-008-9059-x>
- CHEKE, R. A., BASÁÑEZ, M.-G., PERRY, M., WHITE, M. T., GARMS, R., OBUOBIE, E., LAMBERTON, P. H. L., YOUNG, S., OSEI-ATWENEBOANA, M. Y. et al. (2015). Potential effects of warmer worms and vectors on onchocerciasis transmission in West Africa. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370**. <https://doi.org/10.1098/rstb.2013.0559>
- COLEBUNDERS, R., BASÁÑEZ, M.-G., SILING, K., POST, R. J., ROTSAERT, A., MMBANDO, B., SUYKERBUYK, P. and HOPKINS, A. (2018). From river blindness control to elimination: Bridge over troubled water. *Infect. Dis. Poverty* **7** 21. <https://doi.org/10.1186/s40249-018-0406-7>
- CORNUET, J.-M., MARIN, J.-M., MIRA, A. and ROBERT, C. P. (2012). Adaptive multiple importance sampling. *Scand. J. Stat.* **39** 798–812. MR3000850 <https://doi.org/10.1111/j.1467-9469.2011.00756.x>
- DIGGLE, P. and GIORGI, E. (2019). *Model-Based Geostatistics for Global Public Health: Methods and Applications*. CRC Press, London.
- DWYER-LINDGREN, L., CORK, M. A., SLIGAR, A., STEUBEN, K. M., WILSON, K. F., PROVOST, N. R., MAYALA, B. K., VANDER-HEIDE, J. D., COLLISON, M. L. et al. (2019). Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* **570** 189–193.
- ELVIRA, V. and SANTAMARIA, I. (2019). Multiple importance sampling for efficient symbol error rate estimation. *IEEE Signal Process. Lett.* **2** 420–424.
- FOWLER, A. C. and HOLLINGSWORTH, T. D. (2016). The dynamics of *Ascaris lumbricoides* infections. *Bull. Math. Biol.* **78** 815–833. MR3494574 <https://doi.org/10.1007/s11538-016-0164-2>
- FRALEY, C. and RAFTERY, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.

- GIORGI, E., DIGGLE, P. J., SNOW, R. W. and NOOR, A. M. (2018). Geostatistical methods for disease mapping and visualisation using data from spatio-temporally referenced prevalence surveys. *Int. Stat. Rev.* **86** 571–597. MR3882131 <https://doi.org/10.1111/insr.12268>
- GUYATT, H. L., BUNDY, D. A., MEDLEY, G. F. and 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** 139–143.
- HAMLEY, J. I. D., MILTON, P., WALKER, M. and BASÁÑEZ, 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.* **13** e0007557.
- 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. et al. (2013). Global mapping of infectious disease. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **368** 20120250. <https://doi.org/10.1098/rstb.2012.0250>
- HIJMANS, R. J. (2019). Raster: Geographic data analysis and modeling. R package version 2.4-15. <http://CRAN.R-project.org/package=raster>.
- 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** S237–S244.
- HOUGARD, J. M., ALLEY, E. S., YAMÉOGO, L., DADZIE, K. Y. and BOATIN, B. A. (2001). Eliminating onchocerciasis after 14 years of vector control: A proved strategy. *J. Infect. Dis.* **184** 497–503. <https://doi.org/10.1086/322789>
- JACOB, B. G., LOUM, D., LAKWO, T. L., KATHOLI, C. R., HABOMUGISHA, P., BYAMUKAMA, E., TUKAHEBWA, E., CUPP, E. W. and 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. <https://doi.org/10.1371/journal.pntd.0006702>
- KARAGIANNIS-VOULES, D.-A., BIEDERMANN, P., EKPO, U. F., GARBA, A., LANGER, E., MATHIEU, E., MIDZI, N., MWINZI, P., POLDERMAN, A. M. et al. (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. [https://doi.org/10.1016/S1473-3099\(14\)71004-7](https://doi.org/10.1016/S1473-3099(14)71004-7)
- KEELING, M. J. and ROHANI, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ. MR2354763
- KISH, L. (1965). *Survey Sampling*, 1st ed. ed. Wiley, New York.
- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 <https://doi.org/10.1111/j.1467-9868.2011.00777.x>
- MARIN, J.-M., PUDLO, P. and SEDKI, M. (2019). Consistency of adaptive importance sampling and recycling schemes. *Bernoulli* **25** 1977–1998. MR3961237 <https://doi.org/10.3150/18-BEJ1042>
- MARTINO, L., ELVIRA, V., LUENGO, D. and CORANDER, J. (2017). Layered adaptive importance sampling. *Stat. Comput.* **27** 599–623. MR3613588 <https://doi.org/10.1007/s11222-016-9642-5>
- MICHAEL, E., SMITH, M. E., KATABARWA, M. N., BYAMUKAMA, E., GRISWOLD, E., HABOMUGISHA, P., LAKWO, T., TUKAHEBWA, E., MIRI, E. S. et al. (2018). Substantiating freedom from parasitic infection by combining transmission model predictions with disease surveys. *Nat. Commun.* **9** 4324.
- MOSSER, J. F., GAGNE-MAYNARD, W., RAO, P. C., OSGOOD-ZIMMERMAN, A., FULLMAN, N., GRAETZ, N., BURSTEIN, R., UPDIKE, R. L., LIU, P. Y. et al. (2019). Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000–2016: A spatial and temporal modelling study. *Lancet* **4** 1843–1855.
- NORHAYATI, M., FATMAH, M. S., YUSOF, S. and EDARIAH, A. B. (2003). Intestinal parasitic infections in man: A review. *Med. J. Malays.* **58** 296–305.
- 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. et al. (2018). Drug use in the management of uncomplicated malaria in public health facilities in the Democratic Republic of the Congo. *Malar. J.* **17** 189.
- O'HANLON, S. J., SLATER, H. C., CHEKE, R. A., BOATIN, B. A., COFFENG, L. E., PION, S. D. S., BOUSSINESQ, M., ZOURE, H. G. M., STOLK, W. A. et al. (2016). Model-based geostatistical mapping of the prevalence of *Onchocerca volvulus* in West Africa. *PLoS Negl. Trop. Dis.* **10** e0004328.
- PIGOTT, D. M., BHATT, S., GOLDING, N., DUDA, K. A., BATTLE, K. E., BRADY, O. J., MESSINA, J. P., BALARD, Y., BASTIEN, P. et al. (2014). Global distribution maps of the leishmaniases. *eLife* **3** e02851.
- PLAISIER, A. P., VAN OORTMARSSEN, G. J., REMME, J. and HABBEMA, J. P. (1991). The reproductive lifespan of *Onchocerca volvulus* in West African savanna. *Acta Trop.* **48** 271–284.
- POOLE, D. and RAFTERY, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *J. Amer. Statist. Assoc.* **95** 1244–1255. MR1804247 <https://doi.org/10.2307/2669764>
- PULLAN, R. L., SMITH, J. L., JASRASARIA, R. and BROOKER, S. J. (2014). Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasit. Vectors BioMed. Central.* **7** 37.

- RAFTERY, A. E. and BAO, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics* **66** 1162–1173. MR2758504 <https://doi.org/10.1111/j.1541-0420.2010.01399.x>
- RETKUTE, R., TOULOUPOU, P., BASÁÑEZ, M.-G., HOLLINGSWORTH, T. D. and SPENCER, S. E. (2021). Supplement to “Integrating geostatistical maps and infectious disease transmission models using adaptive multiple importance sampling.” <https://doi.org/10.1214/21-AOAS1486SUPPA>, <https://doi.org/10.1214/21-AOAS1486SUPPB>
- RIPLEY, B. D. (1987). *Stochastic Simulation. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR0875224 <https://doi.org/10.1002/9780470316726>
- ROUTLEDGE, I., WALKER, M., CHEKE, R. A., BHATT, S., NKOT, P. B., MATTHEWS, G. A., BALEGUEL, D., DOBSON, H. M., WILES, T. L. et al. (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. *Parasites Vectors* **11** 316. <https://doi.org/10.1186/s13071-018-2864-y>
- RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 <https://doi.org/10.1111/j.1467-9868.2008.00700.x>
- SBERT, M. and HAVRAN, V. (2017). Adaptive multiple importance sampling for general functions. *Vis. Comput.* **33** 6–8.
- SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289–317.
- SIMARRO, P. P., CECCHI, G., FRANCO, J. R., PAONE, M., DIARRA, A., RUIZ-POSTIGO, J. A., FÈVRE, E. M., MATTIOLI, R. C. and JANNIN, J. G. (2012). Estimating and mapping the population at risk of sleeping sickness. *PLoS Negl. Trop. Dis.* **6** e1859. <https://doi.org/10.1371/journal.pntd.0001859>
- SIREN, J., MARTTINEN, P. and CORANDER, J. (2010). Reconstructing population histories from single-nucleotide polymorphism data. *Mol. Biol. Evol.* **28** 673–683.
- TOULOUPOU, P., RETKUTE, R., HOLLINGSWORTH, T. D. and SPENCER, S. E. F. (2020). Statistical methods for linking geostatistical maps and transmission models: Application to lymphatic filariasis in East Africa. *Spat. Spatio-Tempor. Epidemiol.* **100391** 1–10.
- VEACH, E. and GUIBAS, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH* **95** 419–428.
- VERVER, S., WALKER, M., KIM, Y. E., FOBI, G., TEKLE, A. H., ZOURE, H. G. M., WANJI, S., BOAKYE, D. A., KUESEL, A. C. et al. (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** S267–S274.
- WALKER, M., STOLK, W. A., DIXON, M. A., BOTTOMLEY, C., DIAWARA, L., TRAORÉ, M. O., DE VLAS, S. J. and BASÁÑEZ, M. G. (2017). Modelling the elimination of river blindness using long-term epidemiological and programmatic data from Mali and Senegal. *Epidemics* **18** 4–15. <https://doi.org/10.1016/j.epidem.2017.02.005>