Machine learning approaches to calibrate individual-based infectious disease models

Authors: Theresa Reiker[1,2], Monica Golumbeanu[1,2], Andrew Shattock[1,2], Lydia Burgert[1,2], Thomas A. Smith[1,2], Sarah Filippi[3], Ewan Cameron[4,5,6], Melissa A. Penny[1,2*].

Affiliations:

[1]Swiss Tropical and Public Health Institute, Basel, Switzerland.

[2]University of Basel, Basel, Switzerland.

[3] Imperial College London, UK.

[4] Malaria Atlas Project, Big Data Institute, University of Oxford, Oxford, UK.

[5] Curtin University, Perth, Australia.

[6] Telethon Kids Institute, Perth Children's Hospital, Perth, Australia.

*Correspondence to: melissa.penny@unibas.ch

Abstract: Individual-based models have become important tools in the global battle against infectious diseases, yet model complexity can make calibration to biological and epidemiological data challenging. We propose a novel approach to calibrate disease transmission models via a Bayesian optimization framework employing machine learning emulator functions to guide a global search over a multi-objective landscape. We demonstrate our approach by application to an established individual-based model of malaria, optimizing over a high-dimensional parameter space with respect to a portfolio of multiple fitting objectives built from datasets capturing the natural history of malaria transmission and disease progression. Outperforming other calibration methodologies, the new approach quickly reaches an improved final goodness of fit. Per-objective parameter importance and sensitivity diagnostics provided by our approach offer epidemiological insights and enhance trust in predictions through greater interpretability.

One Sentence Summary: We propose a novel, fast, machine learning-based approach to calibrate disease transmission models that outperforms other methodologies

Background

Over the last century, mathematical modelling has become an important tool to analyze and understand disease- and intervention-dynamics for many infectious diseases. Individual-based models (IBMs), where each person is simulated as an autonomous agent, are now widely used. These mathematical models capture heterogeneous characteristics and behaviors of individuals, and are often stochastic in nature. This bottom-up approach of simulating individuals and transmission events enables detailed, robust and realistic predictions on population epidemic trajectories as well as the impact of interventions such as vaccines or new drugs (*1, 2*). Going beyond simpler (compartmental) models to capture stochasticity and heterogeneity in populations, disease progression, and transmission, IBMs can additionally account for contact networks, individual care seeking behavior, immunity effects, or within-human dynamics (*1-3*). As such, well-developed IBMs provide opportunities for experimentation under relatively naturalistic conditions without expensive clinical or population studies. Prominent recent examples of the use of IBMs include assessing the benefit of travel restrictions during the Ebola outbreak 2014–2016 (*4*) and guiding the public health response to the Covid-19 pandemic in multiple countries (*5*). IBMs have also been applied to tuberculosis (*6*), influenza (*7*), dengue, and many other infectious diseases (*2*). Within the field of malaria, several IBMs have been developed over the last 15 years and have been used to support understanding disease and mosquito dynamics (*8-10*), predict the public health impact or carry out economic analyses of (new) interventions (*11-14*); and investigate drug resistance (*15*). Many have had wide-reaching impact, influencing WHO policy recommendations (*11, 16-18*) or strategies of national malaria control programs (*19*).

Calibration caveats and the curse of dimensionality

For model predictions to be meaningful, modelers need to ensure their models accurately capture abstractions of the real world. The potential complexity and realism of IBMs often come at the cost of long simulation times and potentially large numbers of input parameters, whose exact values are often unknown. Parameters may be unknown because they represent derived mathematical quantities that cannot be directly measured or require elaborate, costly experiments (for example shape parameters in decay functions (*20*)), because the data required to derive them in isolation is incomplete or accompanied by inherent biases, or because they interact with other parameters.

Calibrating IBMs poses a complex high-dimensional optimization problem and thus algorithm-based calibration is required to find a parameter set that ensures realistic model behavior, capturing the biological and epidemiological relationships of interest. Local optima may exist in the potentially highly irregular, high-dimensional goodness-of-fit surface, making iterative, purely sampling-based algorithms (e.g. Particle Swarm Optimization or extensions of Newton-Raphson) inefficient and, in light of finite runtimes and computational resources, unlikely to find global optima. Additionally, the *curse of dimensionality* means the number of evaluations of the model scales exponentially with the number of dimensions (*21*). As an example, for the model discussed in this paper, a 23-dimensional parameter space at a sampling resolution of one sample per 10 percentile cell in each dimension, this would yield $10^{number\ of\ dimensions} = 10^{23}$ cells. This is larger than number of stars in the observable Universe (of order 10^{22} (*22*)). Furthermore, most calibrations are not towards one objective or dataset. For multi-objective fitting, each parameter set requires the evaluation of multiple outputs and thus multiple simulations to ensure that all outcomes of interest are captured (in the model discussed here epidemiological outcomes such as prevalence, incidence, or mortality patterns).

In this study, we developed and applied a new approach to calibrate a well-established and used IBM of malaria dynamics called *OpenMalaria*. OpenMalaria features within-host parasite dynamics, the progression of clinical disease, development of immunity, individual care seeking behavior, vector dynamics and pharmaceutical and non-pharmaceutical antimalarial interventions at vector and human level (https://github.com/SwissTPH/openmalaria/wiki) (*3, 20, 23*). Previously, the model was calibrated using an asynchronous genetic algorithm (GA) to fit 23 parameters to 11 objectives representing different epidemiological outcomes, including age-specific prevalence and incidence patterns, age-specific mortality rates and hospitalization rates (*3, 20, 23*) (see supplementary texts 1 and 2 for details on the calibration objectives and data). Genetic algorithms (GAs) build on principles of population genetics and evolution (*selection, reproduction* and *mutation*), evolving a random starting population of candidate parameter sets with each iteration towards the nearest local optimum within the goodness of fit (*fitness*) landscape (*24*). However, the sampling-based nature and sequential function evaluations of GAs can be too slow for high-dimensional problems in irregular spaces where only a limited number of function evaluations are possible. Additionally, similar to the evolution of a population in nature, step-by-

step selection of the fittest individuals evolves the population towards the *nearest* fitness peak making valleys of neutral or lower fitness difficult to cross (*25*).

Other solutions to fit similarly detailed IBMs of malaria employ a combination of directly extracting parameter values from the literature where information is available, and fitting the remainder using multi-stage, modular Bayesian Markov Chain Monte Carlo (MCMC)-based methods (*26-31*). For these models, multiple fitting objectives are often not addressed simultaneously. Rather, to our knowledge, most other malaria IBMs are divided into functional modules (such as the human transmissibility model, within-host parasite dynamics model, and the mosquito or vector model), which are assumed to be influenced by only a limited number of parameters each. The modules are then fit independently and in a sequential manner (*27-31*). Modular approaches reduce the dimensionality of the problem, allowing for the use of relatively straightforward MCMC algorithms. However, these struggle with efficiency in high dimensions as their Markovian nature requires many sequential function evaluations (10^4–10^7 even for simple models), driving up computing time and computational requirements (*32*). Additionally, whilst allowing for the generation of posterior probability distributions of the parameters (*30*), the modular nature makes sequential approaches generally unable to account for interdependencies between parameters assigned to different modules and how their co-variation may affect disease dynamics.

Emulators and Bayesian Optimization

Progress in recent years on numerical methods for supervised, regularized learning of smooth functions from discrete training data allows us to revisit calibration of detailed mathematical models. In particular, Bayesian optimization with Gaussian processes has gained popularity as an efficient approach to tackle expensive optimization problems, for example in hyperparameter search problems in machine learning (*33, 34*). Assuming that the parameter-solution space exhibits a modest degree of regularity, a prior distribution is defined over a computationally expensive objective function by the means of a light-weight probabilistic emulator such as a Gaussian process. The constructed emulator is sequentially refined by adaptively sampling the next training points based on acquisition functions derived from the posterior distribution. The trained emulator model is used to make predictions over the objective functions from the input space with minimum evaluation of the expensive *true* (simulator) function. Purely sampling-based iterative approaches

(like genetic algorithms) are usually limited to drawing sparse random samples from proposals located nearby existing samples in the parameter space. In contrast, the use of predictive emulators permits exploration of the entire parameter space at higher resolution. This increases the chances of finding the true global optimum of the complex objective function in question and avoiding local optima.

Here, we introduce a novel approach for calibrating IBMs (Fig. 1) based on Bayesian optimization and incorporating machine learning algorithms. We prove the strength and versatility of our approach by calibrating the OpenMalaria model and optimizing its 23 input parameters using real-world data on 11 epidemiological outcomes in parallel. To emulate the solution space, we explore and compare two prior distributions, namely a Gaussian process (GP) emulator and a *superlearning* algorithm in form of a Gaussian process stacked generalization (GPSG) emulator. We first use a Gaussian process (GP) emulator to emulate the solution space. Whilst GP emulators provide flexibility whilst retaining relative simplicity (*34*) and have been used previously as priors in Bayesian optimization (*33*), stacked generalization algorithms have not. They provide a potentially attractive alternative as they have been shown to outperform GPs and other machine learning algorithms in capturing complex spaces (*13, 35*). The stacked generalization algorithm (*35*) builds on the idea of creating ensemble predictions from multiple learning algorithms (*level 0 learners*). The cross-validated predictions of the level 0 learners are incorporated into a general learning system (level 1 meta-learner). This allows for the combination of memory-efficient and probabilistic algorithms in order to reduce computational time, whilst retaining probabilistic elements required for adaptive sampling. We prove the superior performance and speed of the Bayesian optimization calibration scheme to classical genetic algorithm approaches and thus propose a novel modus operandi to parameterize complex mathematical models.

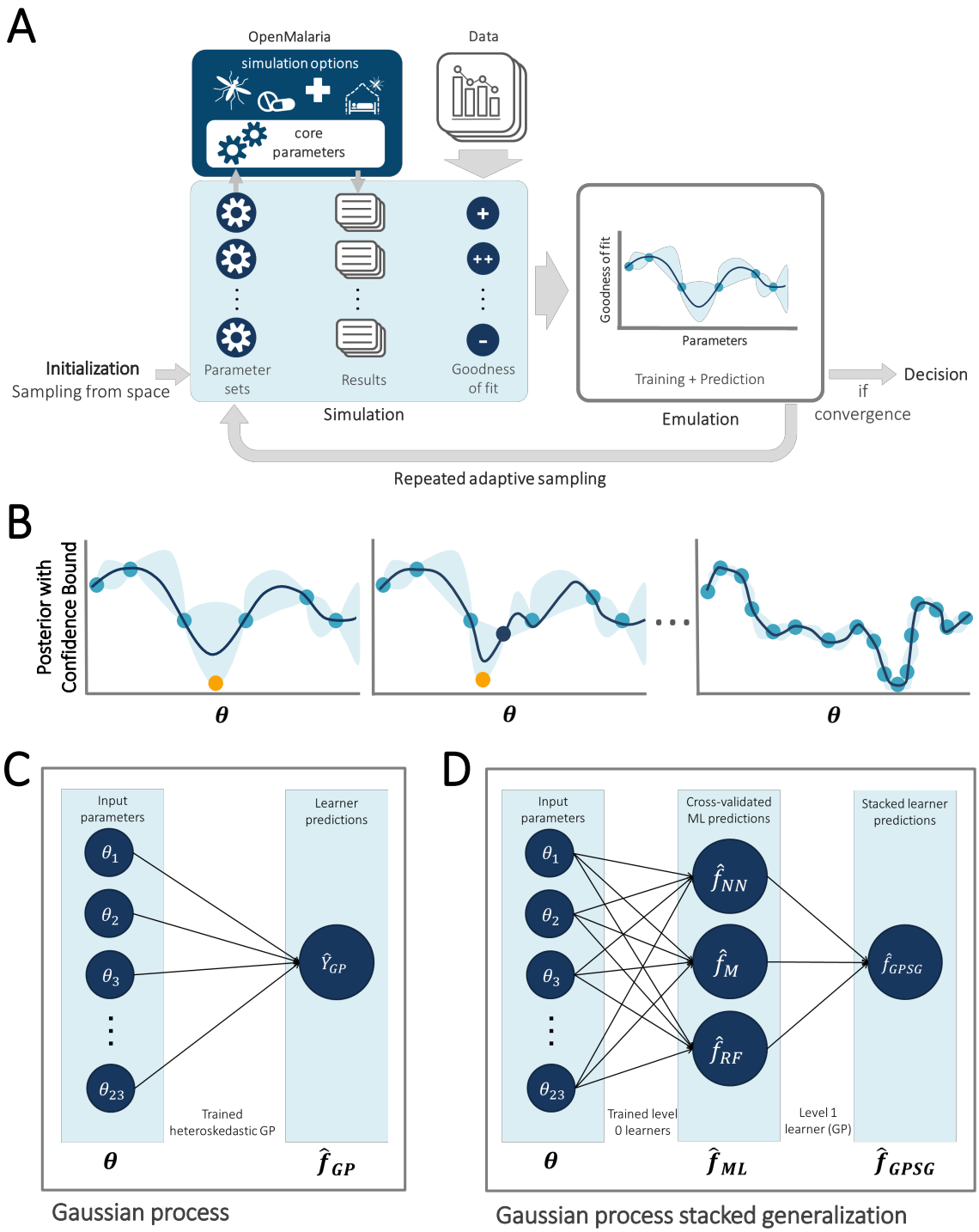


Fig. 1. *Overview of model calibration approaches by Bayesian optimization using Gaussian process and machine learning emulators.* ***A. General Framework****. The input parameter space is initially sampled in a space-filling manner, generating the initial core parameter sets (initialization). For each candidate set, simulations are performed with the model, mirroring the studies that yielded the calibration data. The deviation between simulation results and data is assessed, yielding goodness of fit scores for each parameter set. An emulator (C or D) is trained to capture the relationship between parameter sets and goodness of fit and used to generate out-of-sample predictions. Based on these, the most promising additional parameter sets are chosen (adaptive sampling*

by means of an acquisition function), evaluated, and added to the training set of simulations. Training and adaptive sampling are repeated until the emulator converges and a decision on the parameter set yielding the best fit is made. ***B. Acquisition Function****. The acquisition function is used to determine new parameter space locations. Thus,* $\boldsymbol{\theta}$ *is a vector of input parameters (23-dimensional for the model described here) to be evaluated during adaptive sampling. It incorporates both predictive uncertainty of the emulator and proximity to the minimum.* ***C. Gaussian process emulator.*** *A heteroscedastic Gaussian process is used to generate predictions on the loss functions,* $\hat{\boldsymbol{f}}_{\boldsymbol{GP}}(\boldsymbol{\theta})$***,*** *for each input parameter set* $\boldsymbol{\theta}$*.* ***D. Gaussian process stacked generalization emulator.*** *Three machine learning algorithms (level 0 learners: bilayer neural net, multivariate adaptive regression splines and random forest) are used to generate predictions on the individual objective loss functions* $\hat{\boldsymbol{f}}_{NN}, \hat{\boldsymbol{f}}_{M}$ *and* $\hat{\boldsymbol{f}}_{RF}$ *(collectively* $\hat{\boldsymbol{f}}_{ML}$*) at locations* $\boldsymbol{\theta}$*. These predictions are inputs to a heteroscedastic (level 1 learner) which is used to generate the stacked learner predictions* $\hat{\boldsymbol{f}}_{\boldsymbol{GPSG}}$ *and derive predictions on the overall goodness of fit* $\hat{F}_{GPSG}$*.*

Calibration workflow

The developed model calibration workflow approach is summarized in Figure 1A. In brief, goodness of fit scores were first derived for randomly generated, initial parameter sets. The goodness of fit scores were defined as a weighted sum of the loss functions for each of 11 fitting objectives. These span various epidemiological measures capturing the complexity and heterogeneity of the malaria transmission dynamics, including the age-prevalence and age-incidence relationships, and are informed by a multitude of observational studies (see methods and supplementary text 2). Next, GP and GPSG emulators were trained on the obtained set of scores and used to approximate the relationship between parameter sets and goodness of fit for each objective. After initial investigation of different machine learning algorithms, the GPSG was constructed using a bilayer neural net, multivariate adaptive regression splines and random forest as *level 0* learners and a heteroscedastic Gaussian process as *level 1* learner (Figure 1C-D, see methods and supplement). Using a lower confidence bound acquisition function based on the emulators' point and uncertainty predictions for proposed new candidate parameter sets, the most promising sets were chosen. These parameter sets were simulated and added to the database of simulations for the next iteration of the algorithm. At the next iteration, the emulators are re-trained on the new simulation database and re-evaluated (Figure 1B). This iterative process of simulation, training and emulation was repeated until a memory limit of 1024GB was hit.

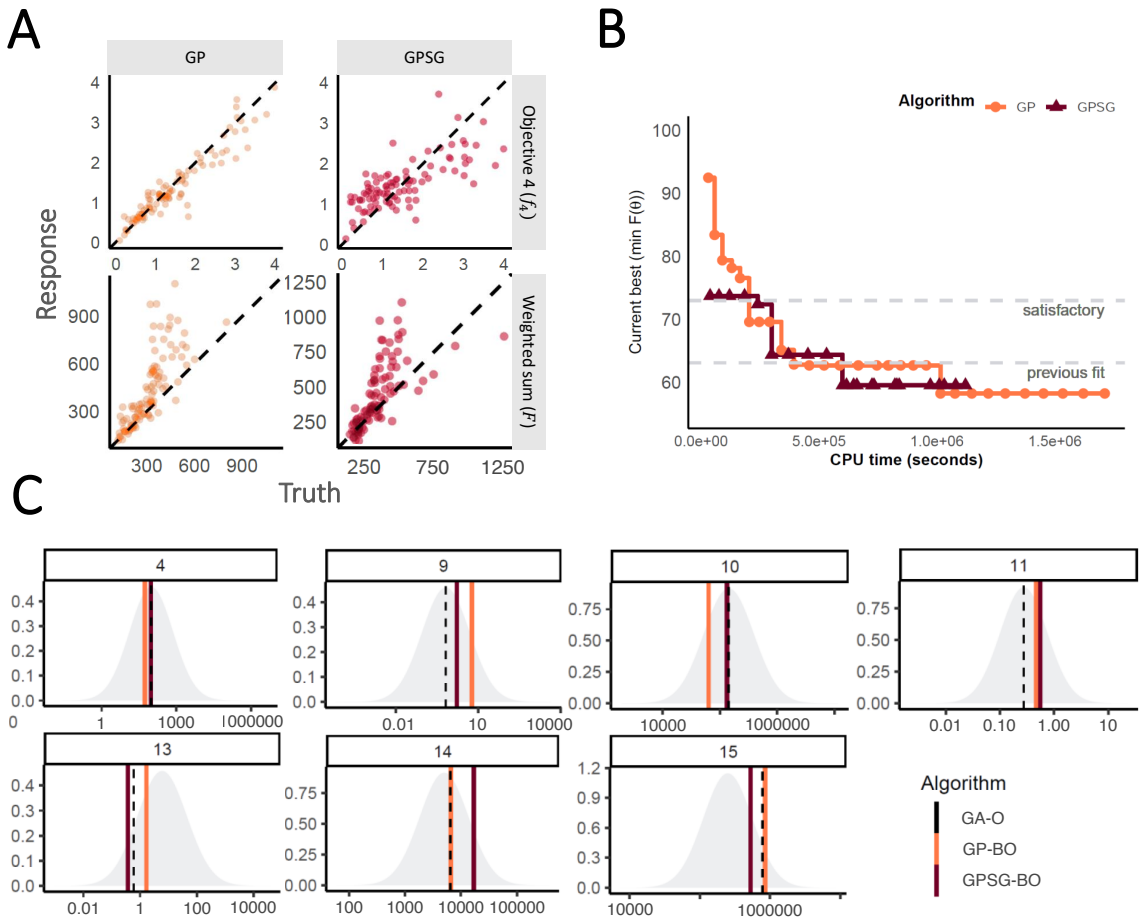


Fig. 2. Emulator performance. A. Example of emulator predictions vs true values on a 10% holdout set*. Predictions are shown for the final iteration of each optimization (iteration 30 for GP-BO and iteration 23 for GPSF-BO). Here, emulator performances are shown for objective 4 (the age-dependent multiplicity of infection,* f_4*) and the weighted sum* F*. Plots for all other objectives are provided in the supplement. GP = Gaussian process emulator, GPSG = Gaussian process stacked generalization emulator* ***B. Convergence.*** *Weighted sum of loss functions over 11 objectives associated with the current best fit parameter set by CPU time in seconds. Satisfactory fit of OpenMalaria refers to a weighted sum of loss functions value of 73.2 (as defined by Smith 2012). Previous best fit for OpenMalaria was achieved by the genetic algorithm had a loss function value of 63.7. Our new approach yields a fit of 58.2 for GP-BO in iteration 21 within in* $1.02e^6$ *CPU seconds (~12 days) and 59.6 for GPSG-BO in iteration 10 in* $6.00e^5$ *CPU seconds (~7 days). GP-BO = Gaussian process emulator Bayesian optimization, GPSG-BO = Gaussian process stacked generalization emulator Bayesian optimization).* ***C. Example log prior parameter distributions and posterior estimates****. The most influential parameters on the weighted sum of the loss functions are shown here (see figure 3C). All other plots can be found in the supplement. The posterior estimates for GP-BO and GPSG-BO are shown in relation to those previously derived through optimization using a genetic algorithm (GA-O)*

Algorithm performance by iteration and time and convergence

Both emulators adequately captured the input-output relationship of the calculated loss-functions from the simulator, with better accuracy when close to minimal values of the weighted sum of the

loss functions, F (Fig. 2A). This is sufficient as the aim of both emulators within the Bayesian optimization framework is to find minimal loss function values rather than an overall optimal predictive performance for all outcome values. Examples of truth vs predicted estimates on a 10% holdout set are provided in Figure 2A (additional plots for all objectives can be found in supplementary figures S2-S5). A *satisfactory fit* of the simulator was previously defined by a loss function value of $F = 73.2$ (*20*). The *previous best* model fit derived using the GA had a weighted sum of the loss functions of $F = 63.7$ (*20*). *Satisfactory fit* was achieved by our approach in the first iteration of the GPSG-based Bayesian optimization algorithm (GPSG-BO), and after six iterations for the GP-based algorithm (GP-BO) (Fig. 2B). The *current best* fit was approximately retrieved after six iterations for the GPSG-BO algorithm and after nine iterations for GP-BO, and was improved by both algorithms after ten iterations (returning final values $F = 58.3$ for GP-BO and 59.6 for GPSG-BO). This shows that the Bayesian optimization approach with either of our emulators very quickly achieves a better simulator fit than obtained with a classical GA approach that was previously employed to calibrate OpenMalaria. Of the two emulators, the GP approach finds a parameter set associated with a better overall accuracy and the GPSG reaches *satisfactory* values faster (both in terms of iterations and CPU time). A likely explanation for this is that the GPSG-BO is unable to propagate its full predictive variance into the acquisition function. Only uncertainty stemming from the level 1 probabilistic learner (GP) is therefore captured in the final prediction. This leads to underestimation of the full predictive variance, and a bias towards exploitation in the early stages of the GPSG-BO algorithm (as illustrated by early narrow sampling, see supplementary figures S6-7).

Figure 2C shows examples of the posterior estimates returned by the optimization algorithms in context of the log prior distributions for the parameters with the greatest effects on F (see also figure 3C). All algorithms return parameter values within the same range and (apart from parameter 4), clearly distinct from the prior mean. The fact that highly similar parameter values are identified by multiple algorithms strengthens confidence in the final parameter sets yielded by the algorithms.

Optimal Goodness of Fit

The best fit parameter sets yielded by our approach are provided in the supplement (Table S2). Importantly, after ten iterations of the GPSG-BO algorithm (approximately 7 days), and 20

iterations for the GP-BO algorithm (approximately 12 days), both approaches yielded similar values of the 11 objective loss functions, along with similar weighted total loss function values, and qualitatively similar visual fits and predicted trends to the data (Fig. 3A-B and supplement). We found this to be an unexpectedly fast result of the two algorithms. Details of the algorithm's best fits to the disease and epidemiological data are shown in Figures S8-S18. Overall, several objectives had visual and reduced loss-function improvements, for example to the objective on the multiplicity of infection (Fig. 3A).

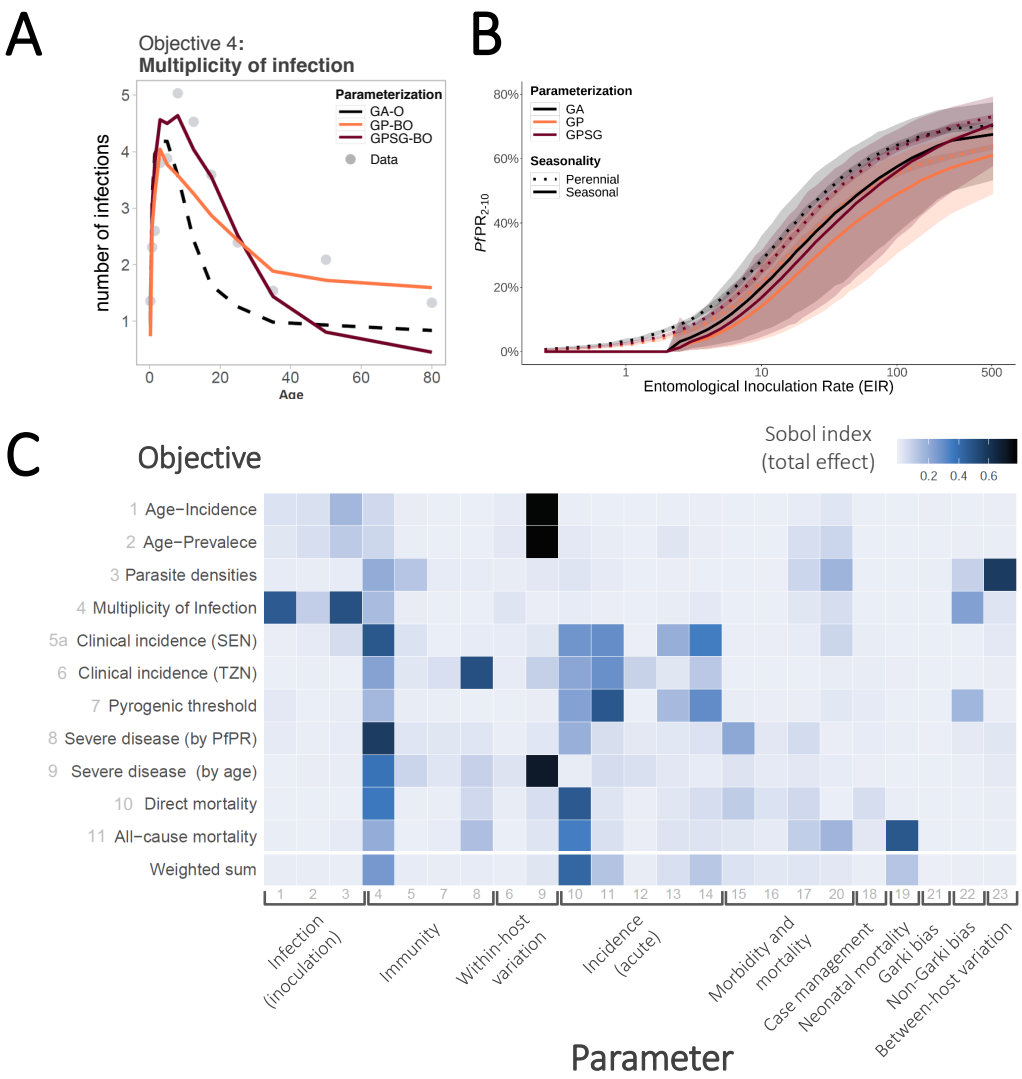


Fig. 3. A. Multiplicity of infection by age. *Comparison of simulator goodness of fit for objective 4, the age-specific multiplicity of infection (number of genetically distinct parasite strains concurrently present in one host). Simulations were carried out for the same random seed for all parameterizations and for a population size of N=5,000.* ***B. Simulated epidemiological relationship between transmission intensity (entomological inoculation rate, EIR) and P. falciparum prevalence (PfPR$_{2\text{-}10}$).*** *Simulated epidemiological relationship between the transmission intensity (EIR in number of infectious bites per person per year) and*

infection prevalence in individuals aged 2-10 years ($PfPR_{2-10}$) under the parameterizations achieved by the different optimization algorithms. Lines show the mean across 100 random seed simulations for a simulated population size N=10,000 and the shaded area shows the 95% confidence interval. ***C. Parameter effects on the objective variance.*** *Using the GP emulator, a global sensitivity analysis (Sobol analysis) was conducted. The tile shading shows the total effect indices for all objective functions and parameters grouped by function. SEN= Senegal, TZN = Tanzania.*

Impact / Parameter sensitivity analysis & External validation

each of the 11 fitting objectives, we used the GP emulator trained on all available training simulation results from the optimization process (R^2=0.53 [objective 7] - 0.92 [objective 3]) to conduct a global sensitivity analysis by variance decomposition (here via Sobol analysis (*36*)). Figure 3C shows Sobol total effect indices quantifying the importance of individual parameters and describing each parameter's contributions to the outcome variance for each objective. Our results indicate that most objectives are influenced by multiple parameters from different groups, albeit to varying degrees, thus highlighting the importance of simultaneous multi-objective fitting. Clusters of influential parameters can be observed for most objectives; for example, parameters associated with incidence of acute disease influence clinical incidence and pyrogenic threshold objectives. Some parameters have strong influence on multiple objectives, such as parameter 4, the critical value of cumulative number of infections and influences immunity acquisition; and parameter 10, a factor required to determine the pyrogenic threshold, which we find to be a key parameter determining infections progressing to clinical illness.

Algorithm validation

In order to test if our algorithms can recover a known solution, the final parameter sets for both approaches were used to generate synthetic field data sets, and our approaches were subsequently applied to recover the known parameter set. For the GP, 13 of the 23 parameters were recovered (Figure S19A). Those not recovered largely represented parameters to which the weighted loss function was found to be insensitive (Figure 3C). Thus, rather than showing a shortcoming of the calibration algorithm, this suggests a potential for dimensionality reduction of the simulator and re-evaluation of its structure. However, this was beyond the scope of the current work to develop a fast and powerful calibration methodology for IBMs.

Comparison of key epidemiological relationships and implications for predictions

The new parameterizations for OpenMalaria were further explored to assess key epidemiological relationships, in an approach similar multiple- model comparison in Penny et al. 2016 (*11*). We

examined incidence and prevalence of disease, as well as incidence of mortality for multiple archetypical settings, considering a range of perennial and seasonal transmission intensity and patterns. The results are presented in Figure 3B and Figures S20-30. The new parameterizations result in increased predicted incidence of severe episodes and decreased prevalence for all transmission intensities (thus also slightly modifying the prevalence-incidence relationship). While we found that the overall implications for the other simulated epidemiological relationship were small, the differences in predictions for severe disease may carry important implications for public health decision making. We conclude that our new parameterizations do not fundamentally bring into question previous research conducted using OpenMalaria, but we do suggest re-evaluation of adverse downstream events such as severe disease and mortality.

Discussion

Calibrating individual-based models can be challenging as many techniques struggle with high dimensionality, or become infeasible with long model simulation times and multiple calibration objectives. However, ensuring adequate model fit to key data is vital, as this impacts the weighting we should give model predictions in the public health decision making process. Our machine learning and Bayesian optimization approaches provide fast solutions to calibrating individual-based models while improving model accuracy, and by extension prediction accuracy.

Using our novel Bayesian optimization approach, we calibrated a detailed simulator of malaria transmission and epidemiology dynamics with 23 input parameters simultaneously to 11 epidemiological outcomes, including age-incidence and -prevalence patterns. The use of a probabilistic emulator to predict goodness-of-fit, rather than conducting sparse sampling, allows for cheap evaluation of the simulator at many locations and increases our confidence that the final parameter set represents a global optimum. Our approach provides a fast calibration whilst also providing a better fit compared to the previous parameterization. We are further able to define formal endpoints to assess calibration alongside *visual confirmation* of goodness of fit (*20, 27*), such as the emulator's predictive variance approaching the observed simulator variance. The emulator's ability to quantify the input stochasticity of the simulator also enables simulation at small population sizes, contributing to fast overall computation times.

Despite the demonstrated strong performance of stacked generalization in other contexts such as geospatial mapping (*13, 35, 37-40*), we found that using a *superlearning* emulator for Bayesian

optimization was not superior to traditional GP-based methods. In our context using GPSG sped up convergence of the algorithm, but both approaches, GP and GPSG, led to equally good fits. Each approach does however, have different properties with context-dependent benefits: The dimensionality reduction provided by GPSG approaches may lead to computational savings depending on the *level 0* and *level 1* learners. At the same time, only level 1 learner uncertainty is propagated into the final` ` predictions, which affects the efficacy of adaptive sampling and may lead to overly exploitative behavior, where sampling close to the point estimate of the predicted optimum is overemphasized, rather than exploring the entire parameter space (see supplements S2 and S3 on selected points). On the other hand, exploration/exploitation trade-offs for traditional GP-BO algorithms have long been examined and *no regret* solutions have been developed (*41*).

Our methodology constitutes a highly flexible framework for individual based model calibration. Both algorithms can be applied to other parameterization and optimization problems in disease modelling and also in other modelling fields, such as physical or mobility and transport models. Furthermore, in the GPSG approach, additional or alternative level 0 can be easily incorporated. Possible extensions to our approach include combination with methods to adaptively reduce the input space for constrained optimization problems (*42*), or other emulators may be chosen depending on the application. For example, homoscedastic GPs, which are faster than the heteroscedastic approach presented here, may be sufficient for many applications (but not for our IBM in which heteroscedastic was required due to the stochastic nature of the model). Alternatively, the computational power required by neural net algorithms scales only linearly (compared with a nominal cubic scaling for GPs) with the sample size, and we envisage wide applications for neural net-based Bayesian optimization in high dimensions. In our example, the bilayer neural net algorithm completed training and prediction within seconds whilst maintaining very high predictive performance. Unfortunately, estimating the uncertainty required for good acquisition functions is difficult in neural networks, but solutions are being developed (*34, 43*). These promising approaches should be explored as they become more widely available in high-level programming languages. With the increased availability of code libraries and algorithms, Bayesian optimization with a range of emulators is also becoming easier to implement.

The probabilistic, emulator-based calibration approach is accompanied by many benefits, including relatively quick global sensitivity analysis. As explored in this work, GP-based methods

are easily coupled with sensitivity analyses, which provide detailed insights into a model's structural dependencies and the sensitivity of its goodness of fit to the input parameters. To the best of our knowledge, no other individual-based model calibration study has addressed this. In the case of malaria models, we have shown the interdependence of all OpenMalaria model components and a relative lack of modularity. In particular, within-host immunity-related parameters were shown to influence all fitting objectives, including downstream events such as severe disease and mortality when an infection progresses to clinical disease. Thus, calibrating within-host immunity in the absence of key epidemiology and population outcomes can lead to suboptimal calibration and ultimate failure of the model to adequately capture disease biology and epidemiology.

We have employed a new approach to calibrating malaria models compared with previous methods, but reach broadly similar comparisons to the natural history of disease. We also attainted a slightly improved but similar goodness of fit, the main benefit being improved fitting times and the ability to measure parameter importance. Given the high number of influential parameters for each epidemiological objective in our parameter importance investigations, and the overlap between parameter-objective associations, we argue that, where possible, multi-objective fitting should be preferred over purely sequential approaches. Our approach confirms that using a parallel approach to parameterization rather than a modular, sequential, one captures the joint effects of all parameters and ensures that all outcomes are simultaneously accounted for. To the best of our knowledge, no model of malaria transmission of comparable complexity and a comparable number of fitting objectives was simultaneously calibrated to all its fitting objectives. Disregarding the joint influence of *all* parameters on the simulated outcomes may negatively impact the accuracy of model predictions, in particular on policy-relevant outcomes of severe disease and mortality.

Despite providing relatively fast calibration towards a better fitting parameter set, several limitations remain in our work. We have not systematically tested that a global optimum has been reached in our new approach, but assume it is close to a global minimum for the current loss-functions defined, as further iterations did not yield changes, and both the GP and GPSG achieved similar weighted loss function and parameter sets. We aimed to improve the algorithm to calibrate detailed IBM, but we did not incorporate new data, which will be important moving forward as our parameter importance and validation analysis highlights several key epidemiological outcomes on severe disease and mortality are sensitive to results.

The key limitations of Bayesian optimization, particularly when using a Gaussian process emulator, are the high computational requirements in terms of memory and parallel computing nodes due to increasing runtimes and cubically scaling memory requirements of GPs. Memory limits may thus be reached before the predictive variance approached its limit. Furthermore, we chose an acquisition function with high probability to be *no regret* (*41*), but this likely overemphasizes exploration in the early stages of the algorithm considering the dimensionality of the problem and finite runtime. We opted here for pure exploitation every 5 iterations, but a more formal optimization of the acquisition function should be explored. The GPSG approach presented here can partially alleviate this challenge, depending on the choice of learning algorithms, but the iterative nature and need for many simulations remain. Memory- and time-saving extensions are thus worth exploring, such as incorporating GPU computing or adaptively constraining the prior parameter space, dimensionality reduction, or addressing alternative acquisition functions. Additionally, as with all calibration methodologies, many choices are left to the user, such as the size of the initial set of simulations, the number of points added per iteration, or the number of replicates simulated at each location. There is no general solution to this as the optimal choices are highly dependent on the problem at hand, and we did not aim to optimize these. Performance might be optimized further through a formal analysis of all these variables, however the methodology here is already fast, effective, and highly generalizable to different types of simulation models and associated optimization problems. Improving the loss-functions or employing alternative *Pareto front* efficiency algorithms was not the focus of our current study but would be a natural extension of our work, as would be alternative approaches to the weighting of objectives, which remains a subjective component of multi-objective optimization problems (*44*).

A model's calibration to known input data forms the backbone of its predictions. Our workflow presented here provides large advances in the calibration of detailed mathematical models of infectious disease.

References and Notes:

1. D. L. DeAngelis, V. Grimm, Individual-based models in ecology after four decades. *F1000prime reports* **6**, (2014).

390 2. L. Willem, F. Verelst, J. Bilcke, N. Hens, P. Beutels, Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). *BMC Infect Dis* **17**, 612 (2017).

3. T. Smith *et al.*, Towards a comprehensive simulation model of malaria epidemiology and control. *Parasitology* **135**, 1507-1516 (2008).

395 4. M. F. Gomes *et al.*, Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLoS Curr* **6**, (2014).

5. N. Ferguson *et al.*, Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. *Imperial College London* **10**, 77482 (2020).

6. T. Cohen *et al.*, Are survey-based estimates of the burden of drug resistant TB too low?
400 Insight from a simulation study. *PLoS One* **3**, e2363 (2008).

7. M. E. Halloran *et al.*, Modeling targeted layered containment of an influenza pandemic in the United States. *Proc Natl Acad Sci U S A* **105**, 4639-4644 (2008).

8. N. Chitnis, D. Hardy, T. Smith, A periodically-forced mathematical model for the seasonal dynamics of malaria in mosquitoes. *Bull Math Biol* **74**, 1098-1124 (2012).

405 9. E. Cameron *et al.*, Defining the relationship between infection prevalence and clinical incidence of Plasmodium falciparum malaria. *Nat Commun* **6**, 8170 (2015).

10. P. A. Eckhoff, Malaria parasite diversity and transmission intensity affect development of parasitological immunity in a mathematical model. *Malar J* **11**, 419 (2012).

11. M. A. Penny *et al.*, Public health impact and cost-effectiveness of the RTS,S/AS01 malaria
410 vaccine: a systematic comparison of predictions from four mathematical models. *Lancet* **387**, 367-375 (2016).

12. H. C. Slater, P. G. Walker, T. Bousema, L. C. Okell, A. C. Ghani, The potential impact of adding ivermectin to a mass treatment intervention to reduce malaria transmission: a modelling study. *J Infect Dis* **210**, 1972-1980 (2014).

415 13. S. Bhatt *et al.*, Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of The Royal Society Interface* **14**, 20170520 (2017).

14. P. Winskill, P. G. Walker, J. T. Griffin, A. C. Ghani, Modelling the cost-effectiveness of introducing the RTS,S malaria vaccine relative to scaling up other malaria interventions in sub-Saharan Africa. *BMJ Glob Health* **2**, e000090 (2017).

15. T. D. Nguyen *et al.*, Optimum population-level use of artemisinin combination therapies: a modelling study. *Lancet Glob Health* **3**, e758-766 (2015).

16. O. J. Brady *et al.*, Role of mass drug administration in elimination of Plasmodium falciparum malaria: a consensus modelling study. *Lancet Glob Health* **5**, e680-e687 (2017).

17. W. H. Organization, Malaria vaccine: WHO position paper–January 2016. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire* **91**, 33-52 (2016).

18. L. Okell *et al.*, in *Malaria Policy Advisory Committee meeting*. (2015), pp. 16-18.

19. M. Runge *et al.*, Simulating the council-specific impact of anti-malaria interventions: a tool to support malaria strategic planning in Tanzania. *PloS one* **15**, e0228469 (2020).

20. T. Smith *et al.*, Ensemble modeling of the likely public health impact of a pre-erythrocytic malaria vaccine. *PLoS Med* **9**, e1001157 (2012).

21. R. E. Bellman, *Dynamic programming*. (Princeton University Press, ed. 6, 1957).

22. A. Craig, in *BBC News*. (2003), vol. 2020.

23. T. Smith *et al.*, Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of Plasmodium falciparum malaria: Overview. *The American journal of tropical medicine and hygiene* **75**, 1-10 (2006).

24. D. E. Goldberg, Genetic algorithms in search. *Optimization, and MachineLearning*, (1989).

25. P. S. Oliveto, T. Paixão, J. Pérez Heredia, D. Sudholt, B. Trubenová, in *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. (2016), pp. 1163-1170.

26. C. M. Hazelbag, J. Dushoff, E. M. Dominic, Z. E. Mthombothi, W. Delva, Calibration of individual-based models to epidemiological data: A systematic review. *PLoS Comput Biol* **16**, e1007893 (2020).

27. P. A. Eckhoff, A malaria transmission-directed model of mosquito life cycle and ecology. *Malaria journal* **10**, 303 (2011).

28. P. Eckhoff, P. falciparum infection durations and infectiousness are shaped by antigenic variation and innate and adaptive host immunity in a mathematical model. *PloS one* **7**, e44950 (2012).

29. P. Eckhoff, Mathematical models of within-host and transmission dynamics to determine effects of malaria interventions in a variety of transmission settings. *The American journal of tropical medicine and hygiene* **88**, 817-827 (2013).

30. J. T. Griffin *et al.*, Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med* **7**, e1000324 (2010).

31. J. T. Griffin, N. M. Ferguson, A. C. Ghani, Estimates of the changing age-burden of Plasmodium falciparum malaria disease in sub-Saharan Africa. *Nature communications* **5**, 1-10 (2014).

32. I. Fer *et al.*, Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences (Online)* **15**, (2018).

33. J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* **25**, 2951-2959 (2012).

34. J. Snoek *et al.*, in *International conference on machine learning*. (2015), pp. 2171-2180.

35. D. H. Wolpert, Stacked generalization. *Neural networks* **5**, 241-259 (1992).

36. I. M. Sobol, Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment* **1**, 407-414 (1993).

37. D. Benkeser, C. Ju, S. Lendle, M. van der Laan, Online cross-validation-based ensemble learning. *Statistics in medicine* **37**, 249-260 (2018).

38. L. Breiman, Stacked regressions. *Machine learning* **24**, 49-64 (1996).

39. M. J. Van der Laan, E. C. Polley, A. E. Hubbard, Super learner. *Statistical applications in genetics and molecular biology* **6**, (2007).

40. J. Sill, G. Takács, L. Mackey, D. Lin, Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*, (2009).

41. N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, (2009).

42. R. Moriconi, M. P. Deisenroth, K. S. Kumar, High-dimensional Bayesian optimization using low-dimensional feature spaces. *Machine Learning* **109**, 1925-1943 (2020).

43. D. Zhou, L. Li, Q. Gu, in *International Conference on Machine Learning*. (PMLR, 2020), pp. 11492-11502.

44. R. T. Marler, J. S. Arora, The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization* **41**, 853-862 (2010).

45. M. Binois, R. B. Gramacy, M. Ludkovski, Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics* **27**, 808-821 (2018).

46. A. Hadji, B. Szábo, Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? *arXiv preprint arXiv:1904.01383*, (2019).

47. F. D. Foresee, M. T. Hagan, in *Proceedings of International Conference on Neural Networks (ICNN'97)*. (IEEE, 1997), vol. 3, pp. 1930-1935.

48. D. J. MacKay, Bayesian interpolation. *Neural computation* **4**, 415-447 (1992).

49. P. Rodriguez, D. Gianola, BRNN: Bayesian regularization for feed-forward neural networks. *R package version 0.6*, (2016).

50. T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics (ed. Corrected printing 2002, 2009), vol. 2, pp. 533 S.

51. L. Breiman, Random forests. *Machine learning* **45**, 5-32 (2001).

52. A. Liaw, M. Wiener, L. Breiman, A. Cutler. (2015).

53. N. Chitnis *et al.*, Theory of reactive interventions in the elimination and control of malaria. *Malaria journal* **18**, 266 (2019).

54. T. Reiker, N. Chitnis, T. Smith, Modelling reactive case detection strategies for interrupting transmission of Plasmodium falciparum malaria. *Malaria journal* **18**, 259 (2019).

55. M.-L. Cauwet *et al.*, in *International Conference on Machine Learning*. (PMLR, 2020), pp. 1338-1348.

56. S. Kucherenko, D. Albrecht, A. Saltelli, Exploring multi-dimensional spaces: A comparison of Latin hypercube and quasi Monte Carlo sampling techniques. *arXiv preprint arXiv:1505.02350*, (2015).

57. E. Brochu, V. M. Cora, N. De Freitas, A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, (2010).

58. B. Bischl *et al.*, mlr: Machine Learning in R. *The Journal of Machine Learning Research* **17**, 5938-5942 (2016).

59. B. Ripley, W. Venables, M. B. Ripley, Package 'nnet'. *R package version* **7**, 3-12 (2016).

60. T. Hastie, J. Qian, Glmnet vignette. *Retrieved June* **9**, 1-30 (2014).

61. B. Hofner, A. Mayr, N. Robinzonov, M. Schmid, Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational statistics* **29**, 3-35 (2014).

62. H. Ishwaran, U. B. Kogalur, M. U. B. Kogalur, Package 'randomForestSRC'. (2020).

63. M. N. Wright, A. Ziegler, ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*, (2015).

64. N. Meinshausen, M. N. Meinshausen, Package 'nodeHarvest'. (2015).

65. M. J. Jansen, Analysis of variance designs for model output. *Computer Physics Communications* **117**, 35-43 (1999).

66. A. Saltelli *et al.*, Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer physics communications* **181**, 259-270 (2010).

67. N. Maire *et al.*, A model for natural immunity to asexual blood stages of Plasmodium falciparum malaria in endemic areas. *The American journal of tropical medicine and hygiene* **75**, 19-31 (2006).

68. W. E. Collins, G. M. Jeffery, A retrospective examination of sporozoite-and trophozoite-induced infections with Plasmodium falciparum: development of parasitologic and clinical immunity during primary infection. *The American journal of tropical medicine and hygiene* **61**, 4-19 (1999).

69. T. Smith *et al.*, Relationship between the entomologic inoculation rate and the force of infection for Plasmodium falciparum malaria. *The American journal of tropical medicine and hygiene* **75**, 11-18 (2006).

70. G. F. Killeen, A. Ross, T. Smith, Infectiousness of malaria-endemic human populations to vectors. *The American journal of tropical medicine and hygiene* **75**, 38-45 (2006).

71. A. Ross, G. Killeen, T. Smith, Relationships between host infectivity to mosquitoes and asexual parasite density in Plasmodium falciparum. *The American journal of tropical medicine and hygiene* **75**, 32-37 (2006).

72. I. A. Carneiro *et al.*, Modeling the relationship between the population prevalence of Plasmodium falciparum malaria and anemia. *The American journal of tropical medicine and hygiene* **75**, 82-89 (2006).

73. A. Ross, T. Smith, The effect of malaria transmission intensity on neonatal mortality in endemic areas. *The American journal of tropical medicine and hygiene* **75**, 74-81 (2006).

74. T. Smith *et al.*, An epidemiologic model of the incidence of acute illness in Plasmodium falciparum malaria. *The American journal of tropical medicine and hygiene* **75**, 56-62 (2006).

75. A. Ross, N. Maire, L. Molineaux, T. Smith, An epidemiologic model of severe morbidity and mortality caused by Plasmodium falciparum. *The American journal of tropical medicine and hygiene* **75**, 63-73 (2006).

76. G. Port, P. Boreham, J. H. Bryan, The relationship of host size to feeding by mosquitoes of the Anopheles gambiae Giles complex (Diptera: Culicidae). *Bulletin of Entomological Research* **70**, 133-144 (1980).

77. R. H. *et al.*, "The epidemiology of severe malaria due to *Plasmodium falciparum* at different transmisison intensities in NE Tanzania," *LSTMH Malaria Centre R 2002-2003* (2004).

78. L. Molineaux, G. Gramiccia, W. H. Organization, *The Garki project: research on the epidemiology and control of malaria in the Sudan savanna of West Africa*. (World Health Organization, 1980).

79. S. Owusu-Agyei, T. Smith, H. P. Beck, L. Amenga-Etego, I. Felger, Molecular epidemiology of Plasmodium falciparum infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana. *Tropical Medicine & International Health* **7**, 421-428 (2002).

80. T. Smith *et al.*, Absence of seasonal variation in malaria parasitaemia in an area of intense seasonal transmission. *Acta tropica* **54**, 55-72 (1993).

81. A. Y. Kitua *et al.*, Plasmodium falciparum malaria in the first year of life in an area of intense and perennial transmission. *Tropical Medicine & International Health* **1**, 475-484 (1996).

82. W. C. Earle, M. Perez, Enumeration of parasites in the blood of malarial patients. *Journal of laboratory and clinical medicine* **17**, (1932).

83. J.-F. Trape, C. Rogier, Combating malaria morbidity and mortality by reducing transmission. *Parasitology today* **12**, 236-240 (1996).

84. J.-F. Trape *et al.*, The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. *The American journal of tropical medicine and hygiene* **51**, 123-137 (1994).

85. J. Charlwood *et al.*, Incidence of Plasmodium falciparum infection in infants in relation to exposure to sporozoite-infected anophelines. *The American journal of tropical medicine and hygiene* **59**, 243-251 (1998).

86. K. Marsh, R. Snow, Malaria transmission and morbidity. *Parassitologia* **41**, 241 (1999).

87. R. W. Snow *et al.*, Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *The lancet* **349**, 1650-1654 (1997).

88. E. L. Korenromp, B. G. Williams, E. Gouws, C. Dye, R. W. Snow, Measurement of trends in childhood malaria mortality in Africa: an assessment of progress toward targets based on verbal autopsy. *The Lancet infectious diseases* **3**, 349-358 (2003).

89. C. Rogier, D. Commenges, J.-F. Trape, Evidence for an age-dependent pyrogenic threshold of Plasmodium falciparum parasitemia in highly endemic populations. *The American journal of tropical medicine and hygiene* **54**, 613-619 (1996).

90. P. Vounatsou, T. Smith, A. Kitua, P. Alonso, M. Tanner, Apparent tolerance of Plasmodium falciparum in infants in a highly endemic area. *Parasitology* **120**, 1-9 (2000).

91. J. C. Beier, G. F. Killeen, J. I. Githure, Entomologic inoculation rates and Plasmodium falciparum malaria prevalence in Africa. *The American journal of tropical medicine and hygiene* **61**, 109-113 (1999).

92. G. Barnish *et al.*, Malaria in a rural area of Sierra Leone. I. Initial results. *Annals of Tropical Medicine & Parasitology* **87**, 125-136 (1993).

93. I. D. R. Centre, I. Network, *Population and Health in Developing Countries: Population, health and survival at INDEPTH sites*. (IDRC, 2002), vol. 1.

94. H. C. Spencer *et al.*, Impact on mortality and fertility of a community-based malaria control programme in Saradidi, Kenya. *Annals of Tropical Medicine & Parasitology* **81**, 36-45 (1987).

95. U. D'Alessandro *et al.*, Mortality and morbidity from malaria in Gambian children after introduction of an impregnated bednet programme. *The Lancet* **345**, 479-483 (1995).

96. P. Duboz, J. Vaugelade, M. Debouverie, "Mortalité dans l'enfance dans la région de Niangoloko," (ORSTOM, Ouagadougou, Burkina Faso, 1989).

97. J. A. Schellenberg *et al.*, KINET: a social marketing programme of treated nets and net treatment for malaria control in Tanzania, with evaluation of child health and long-term survival. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93**, 225-231 (1999).

98. Z. Premji *et al.*, Community based studies on childhood mortality in a malaria holoendemic area on the Tanzanian coast. *Acta tropica* **63**, 101-109 (1997).

99. J.-F. Trape *et al.*, Impact of chloroquine resistance on malaria mortality. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie* **321**, 689-697 (1998).

Acknowledgments: We acknowledge and thank our colleagues in the Swiss TPH Disease Dynamics unit. **Funding:** The work was funded by the Swiss National Science Foundation through SNSF Professorship of MAP (PP00P3_170702) supporting MAP, MG, and LB. TR was supported by Bill & Melinda Gates Foundation Project OPP1032350. EC's research is supported by funding from the Bill and Melinda Gates Foundation to Curtin University (Opportunity ID: OPP1197730).
Author contributions: MAP and EC conceived the study. Algorithm development by EC, TR, MAP, and SF with implementation and preparation for sharing on GitHub by TR. Loss functions by MAP and TAS. Sensitivity analysis by TR with inputs from MG and LB. First draft was written by TR and MAP, all authors contributed to writing and interpretation of results and approved the final manuscript. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** Code is publicly available on GitHub under https://github.com/reikth/BayesOpt_Calibration and all calibration data is available from the researchers on request.

Supplementary Materials:

Materials and Methods

Supplementary Texts 1 and 2

Figures S1-S32

Tables S1-S6

References (*1-99*)

Supplementary Materials for

Machine learning approaches to calibrate individual-based infectious disease models

Theresa Reiker[1,2], Monica Golumbeanu[1,2], Andrew Shattock[1,2], Lydia Burgert[1,2], Thomas A. Smith[1,2], Sarah Filippi[3], Ewan Cameron[4,5,6], Melissa A. Penny[1,2*].

Affiliations:

[1]Swiss Tropical and Public Health Institute, Basel, Switzerland.

[2]University of Basel, Basel, Switzerland.

[3] Imperial College London, UK.

[4] Malaria Atlas Project, Big Data Institute, University of Oxford, Oxford, UK.

[5] Curtin University, Perth, Australia.

[6] Telethon Kids Institute, Perth Children's Hospital, Perth, Australia.

*Correspondence to: melissa.penny@unibas.ch

This PDF file includes:

Materials and Methods

Supplementary Texts 1 and 2

Figs. S1 to S32

Tables S1 to S6

Contents

1 Materials and Methods 3
1.1 Preparation of calibration data and simulation experiments 3
1.2 General Bayesian Optimization framework with emulators 4
1.3 Malaria transmission and disease simulator 4
1.4 Calibrating OpenMalaria: loss functions and general approach 5
1.5 Emulator definition 7
1.5.1 Heteroskedastic Gaussian Process (hetGP). 7
1.5.2 Gaussian Process Stacked Generalization (GPSG). 7
1.6 Emulator performance 8
1.7 Sensitivity analysis 8
1.8 Synthetic data validation 8
1.9 Epidemiological outcome comparison 8
1.10 Software 9
2 Supplementary text 1: Malaria Transmission Model 10
2.1 Main features 10
2.2 Infection of the human host 11
2.2.1 Differential feeding by mosquitoes depending on body surface area 11
2.2.2 Control of pre-erythrocytic stages 11
2.2.3 Course of infection in the human host 12
2.2.4 Infectivity of the human host 13
2.3 Morbidity 14
2.3.1 Acute morbidity (uncomplicated clinical cases) 14
2.3.2 Severe disease 15
2.3.3 Mortality 16
3 Supplementary text 2: Calibration Approach and Data Summary 1
3.1 Objectives: Epidemiological data and loss functions 2
3.1.1 Age pattern of incidence after intervention 2
3.1.2 Age patterns of prevalence 2
3.1.3 Age patterns of parasite density 3
3.1.4 Age pattern of number of concurrent infections 4
3.1.5 Age pattern of incidence of clinical malaria 5
3.1.6 Age pattern of incidence of clinical malaria: infants 6
3.1.7 Age pattern of threshold parasite density for clinical attacks 6
3.1.8 Hospitalization rate in relation to prevalence in children 7
3.1.9 Age pattern of hospitalization: severe malaria 9
3.1.10 Malaria specific mortality in children (< 5 years old) 11

3.1.11 Indirect malaria infant mortality rate 11
3.2 Tables S3-S4 13
4 Emulator performance 21
5 Adaptive sampling: selected points 25
5.1 GP-BO 25
5.2 GPSG-BO 26
6 OpenMalaria: Final simulator fit 27
7 Validation 33
8 OpenMalaria simulated epidemiology 35
9 Log prior distributions 44
10 Ranger importance 45

1 MATERIALS AND METHODS

1.1 Preparation of calibration data and simulation experiments

Disease transmission models generally have two types of parameter inputs: core parameters, inherent to the disease and determining how its natural history is captured, and simulation options characterizing the specific setting and the interventions in place (Figure 1A in the main manuscript). The simulation options specify the simulation context such as population demographics, transmission intensity, seasonality patterns and interventions and typically vary depending on the simulation experiment. In contrast, the core parameters determine how its epidemiology and aetiopathogenesis are captured. These include parameters for the description of immunity (e.g. decay of maternal protection), or for defining clinical severe episodes (e.g. parasitemia threshold). To inform the estimation of core parameters, epidemiological data on the natural history of malaria extracted from published literature and collated in previous calibrations of OpenMalaria (*3, 20, 23*) were re-used in this calibration round. These include demographic data such as age-stratified numbers of host individuals which are used to derive a range of epidemiological outcomes such as age-specific prevalence and incidence patterns, mortality rates and hospitalization rates.

Site-specific OpenMalaria simulations were prepared, representing the studies that yielded these epidemiological data in terms of transmission intensity, seasonal patterns, vector species, intervention history, case management, and diagnostics (*23*). The mirroring of field study characteristics in the simulation options ensured that any deviation between simulation outputs and data could be attributed to the core parameters. Age-stratified simulation outputs to match to the data include numbers of host individuals, patent infections, and administered treatments. A summary of the data is provided in the supplementary text 2.

1.2 General Bayesian Optimization framework with emulators

In our proposed Bayesian optimization framework (Figure 1) we evaluated the deviation between simulation outputs and the epidemiological data by training probabilistic emulator functions that approximate the relationship between core parameter sets and goodness of fit. To test the optimization approach in this study we considered the original goodness of fit metrics for OpenMalaria detailed in (*20*) and in supplementary text 2, which uses either Residual Sum of Squares (RSS) or negative log-likelihood functions depending on the epidemiological data for each objective (*20, 23*). The objective function to be optimized is a weighted sum of the individual objectives' loss functions.

We adopted a Bayesian optimization framework where a probabilistic emulator function is constructed to make predictions over the loss functions for each objective from the input space, with a minimum amount of evaluations of the (computationally expensive) simulator.

We compared two emulation approaches. Firstly, a heteroskedastic Gaussian process (GP) emulator and secondly a stacked generalization emulator (*35*). For approach 1 (*GP-BO*), we fitted a heteroskedastic Gaussian process with the input noise modelled as another Gaussian process (*45*) with a Matérn 5/2 kernel to account for the high variability in the parameter space (Figure 1C) (*33, 46*). For approach 2 (*GPSG-BO*), we selected a two-layer neural network (*47-49*), multivariate adaptive regression splines (*50*), and a random forest algorithm (*51, 52*) as level 0 learners.

With each iteration of the algorithm, the training was extended using adaptive sampling based on an acquisition function (*lower confidence bound*) that accounts for uncertainty and predicted proximity to the optimum of proposed locations (Figure 1B). As the emulator performance converges (as assessed by its predictive performance on the test set) we gain confidence in the currently predicted optimum.

1.3 Malaria transmission and disease simulator

We applied our novel calibration approach to OpenMalaria (https://github.com/SwissTPH/openmalaria.wiki.git), an open source modelling platform of malaria epidemiology and control. It features several related individual-based stochastic models of *P. falciparum* malaria transmission and control. Overall, the OpenMalaria IBM consists of a model of malaria in humans linked to a model of malaria in mosquitoes and accounts for individual level heterogeneity in humans (in exposure, immunity, and clinical progression) as well as aspects of vector ecology (e.g. seasonality and the mosquito feeding cycle). Stochasticity is featured by including between- and within-host stochastic variation in parasite densities with downstream effects on immunity (*23*). OpenMalaria further includes aspects of the health system context (e.g. treatment seeking behavior and standard of care) (*3, 23*) with additional probabilistic elements such as treatment seeking probabilities or the option for stochastic results of diagnostic tests. An ensemble of

OpenMalaria model alternative variants is available defined by different assumptions about immunity decay, within-host dynamics, heterogeneity of transmission, along with more detailed sub-models that track parasite genetics, and pharmacokinetic and pharmacodynamics. The models allow for the simulation of interventions, such as the distribution of insecticide treated nets (ITNs), vaccines, or reactive case detection (*53, 54*), in comparatively realistic settings. Full details of the model and the history of calibration can be found in the original publications (*3, 20, 23*) and are summarized in supplementary texts 1 and 2. In our application, we use the term *simulator* to refer to the OpenMalaria base model variant (*20*).

1.4 Calibrating OpenMalaria: loss functions and general approach

Aim

Let $\boldsymbol{f}(\boldsymbol{\theta})$ denote a vector of loss functions obtained by calculating the goodness of fit between simulation outputs and the real data (full details of loss function can be found in supplementary text 2). In order to ensure a good fit of the model, we aim to find the parameter set $\boldsymbol{\theta}$ that achieves the minimum of the weighted sum of 11 loss functions (corresponding to the 10 fitting objectives) $F(\boldsymbol{\theta}) = \sum_{i=1}^{11} w_i f_i(\boldsymbol{\theta})$, where $f_i(\boldsymbol{\theta})$ is the value of objective function i at $\boldsymbol{\theta}$ and w_i is the weight assigned to objective function i:

$$\underset{\theta}{\operatorname{argmin}}\left(\sum_{i=1}^{11} w_i f_i(\boldsymbol{\theta})\right)$$

The weights are kept consistent with previous rounds of calibration and chosen such that different epidemiological quantities contributed approximately equally to $F(\boldsymbol{\theta})$ (see supplementary text 2).

Step 1: Initialization. Let $D = 23$ denote the number of dimensions of the input parameter space $\boldsymbol{\Theta}$ and $W = 11$ the number of objective functions $f_i(\theta), i = 1, \dots, 11$. Prior distributions consistent with previous fitting runs (*20*) were placed on the input parameters. As each parameter is measured in different units, we sampled from the D-dimensional unit cube $\boldsymbol{\Theta}$ and converted these to quantiles of the prior distributions (*20*) (supplementary text 2and Figure S6). Previous research suggests that in high-dimensional spaces quasi-Monte Carlo (qMC) sampling outperforms random or Latin Hypercube designs for most function types and leads to faster rates of convergence (*55, 56*). We therefore used Sobol sequences to sample 1,000 initial locations from $\boldsymbol{\Theta}$. The GP can account for input stochasticity of the simulator. For each sample, we simulated 2 random seeds at a population size of 10,000 individuals. Additionally, 100 simulations were run at the centroid location of the unit cube to gain information on the simulator noise. Using small noisy simulations with small populations speeds up the fitting as the noisy simulations are less computational expensive than larger population runs. Replicates were used to detect signals in noisy settings and estimate the pure simulation variance (*45*).

Computational savings were later achieved through pre-averaging of replicates (*45*). The 2000 unique locations were randomly split into a training set (90%) and a test set (10%). All simulator realizations at the centroid were added to the training set.

Step 2: Emulation

2.1: Emulator Training

Each emulator type for each objective function was trained in parallel to learn the relationships between the normalized input space $\boldsymbol{\Theta}$, and the log-transform of the objective functions $\boldsymbol{f}(\boldsymbol{\theta})$. In each dimension $d \in D$, the mean μ_d and standard deviation σ_d of the training set were recorded, $d = 1, \dots, 23$.

2.2 Posterior prediction

We randomly sampled 500,000 test locations in $\boldsymbol{\Theta}$ from a multivariate normal distribution with mean $\boldsymbol{\theta_{opt}}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\theta_{opt}}$ is the location of the current best location and $\boldsymbol{\Sigma}$ is determined based on previously all sampled locations, and scaled each dimension to mean μ_d and standard deviation σ_d. The trained emulators were used to make predictions $\widehat{\boldsymbol{F}(\boldsymbol{\theta})}$ of the objective functions $\boldsymbol{F}(\boldsymbol{\theta})$ at the test locations. Mean estimates, standard deviations, and nugget terms were recorded. The full predictive variance at each location $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ corresponds to the sum of the standard deviation and nugget terms. From this, we derived the weighted sum $(\boldsymbol{\theta}) = \sum_{i=1}^{11} w_i f_i(\boldsymbol{\theta})$, using weights $\boldsymbol{w}$ consistent with previous fitting runs (Smith 2012) with greater weighting for further downstream objectives. The predicted weighted loss function at location $\boldsymbol{\theta}$ was denoted $\hat{F}(\boldsymbol{\theta})$ with a predicted mean $\hat{\mu}_F(\boldsymbol{\theta})$ and variance $\hat{\sigma}_F(\boldsymbol{\theta})$. Every 15 iterations, we increase the test location sample size to 5 Million to achieve denser predictions.

Step 3: Acquisition. We chose the lower confidence bound (LCB) acquisition function to guide the search of the global minimum. Lower acquisition corresponds to *potentially* low values of the weighted objective function, either because of a low mean prediction value or large uncertainty (*57*). From the prediction set at iteration t, we sample without replacement **250** new locations $\boldsymbol{\theta} = \mathrm{argmin}_{\boldsymbol{\theta}} \{\hat{\mu}_F(\boldsymbol{\theta}, t) - \sqrt{\nu \tau_t}\ \hat{\sigma}_t(\boldsymbol{\theta}, t)\}$, with the hyperparameter $\nu = 1$ and $\tau_t = 2\log(T_t^{D/2+2}\pi/\ 3\delta)$, where T_t is the number of previous unique realisations of the simulator at iteration t, and $\delta = 0.01$ is a hyperparameter (Srinivas 2010). We choose this method as with high probability it is *no regret* (Brochu 2010, Srinivas 2010). With increasing iterations, confidence bound-based methods naturally transition from mainly exploration to exploitation of the current estimated minimum. In addition to this, we force exploitation every 10 iterations by setting $\tau_t = 0$).

Step 4: Simulate. The simulator was evaluated at locations identified in step 3 and the realisations were added to the training set. Steps 2-4 were run iteratively. The Euclidian distance between locations of current best realisations was recorded.

Step 5: Convergence. Convergence was defined as no improvement in the best realisation, $\mathrm{argmin}_{\boldsymbol{F}}\ \boldsymbol{F}$.

1.5 Emulator definition

We compared two emulation approaches. Firstly, a heteroskedastic GP emulator and secondly a stacked generalization emulator (*35*) using a two-layer neural net, multivariate adaptive regression splines (MARS) and a random forest as level 0 learners and a heteroskedastic GP as level 1 learner:

1.5.1 Heteroskedastic Gaussian Process (hetGP).

We fitted a Gaussian process with the input noise modelled as another Gaussian process (*45*). After initial exploration of different kernels, we chose a Matérn 5/2 kernel to account for the high variability in the parameter space. A Matérn 3/2 correlation function was also tested performed equally. Each time the model was built (for each objective at each iteration), its likelihood was compared to that of a homoscedastic Gaussian process and the latter was chosen if its likelihood was higher. This resulted in a highly flexible approach, choosing the best option for the current task.

1.5.2 Gaussian Process Stacked Generalization (GPSG).

Stacked generalization was first proposed by Wolpert 1992 (*35*) and builds on the idea of creating ensemble predictions from multiple learning algorithms (level 0 learners). In *superlearning*, the cross-validated predictions of the level 0 learners are fed into a level 1 meta-learner. We compared the 10-fold cross-validated predictive performance of twelve machine learning algorithms on the test set. All algorithms were accessed through the mlr package in R (*58*). We compared two neural network algorithms (brnn (*48*) for a two layer neural network and nnet for a single-hidden-layer neural network (*59*)), five regression algorithms (cvglmnet (*60*) for a generalised linear model with LASSO or Elasticnet Regularization and 10-fold cross validated lambda, glmboost (*61*) for a boosted generalized linear model, glmnet (*60*) for a regular GLM with Lasso or Elasticnet Regularisation, mars for multivariate adaptive regression splines, and cubist for rule-and instance-based regression modelling), three random forest algorithms (randomForest (*52*), randomForestSRC (*62*) and ranger (*63*)), and a tree-like node harvesting algorithm (nodeHarvest (*64*)). Extreme gradient boosting and support vector regression were also tested but excluded from the comparison due to its long runtime. Their performance was compared with regards to runtimes, and correlation coefficients between predictions on the test set and the true values. Based on these, we selected the two-layer neural network (brnn), multivariate adaptive regression spline (mars), and random forest (randomForest) algorithms. This ensemble of machine learning models constituted the level 0 learners and was fitted to the initialization set. Out-of-sample predictions from a 10-fold cross validation of each observation

were used to fit the level 1 heteroskedastic Gaussian process. As in approach 1, we opted for a Matérn 5/2 kernel and retained the option of changing to a homoscedastic model where necessary.

1.6 Emulator performance

We ascertained that both emulators captured the input-output relationship of the simulator by tracking the correlation between true values $\boldsymbol{f}$ and predicted values $\hat{\boldsymbol{f}}$ on the holdout set of 10% of initial simulations with each iteration (truth vs predicted R2 0.51-0.89 for GP vs 0.37-0.77 for GPSG after initialization, see supplement S1). Transition from exploration to exploitation during adaptive sampling was tracked by recording the distribution of points selected during adaptive sampling in each iteration (Figures S2 and S3).

1.7 Sensitivity analysis

A global sensitivity analysis was conducted on a heteroskedastic GP model with Matérn 5/2 kernel that was trained on all training simulation outputs (n=5,400) from the fitting process. We used the Jansen method of Monte Carlo estimation of Sobol' sensitivity indices for variance decomposition (*65, 66*) with 20 000 sample points and 1000 bootstrap replicates. Sobol' indices were calculated for all loss functions $\boldsymbol{f}$ as well as for their weighted sum $\boldsymbol{F}$ and in all dimensions. Whilst keeping the number of sample points to as low as possible for computational reasons, we ascertained that first-order indices summed to 1 and total effects >1. We further ensured that the overall results of the Sobol' analysis were consistent with the results of other global sensitivity analyses, namely the relative parameter importance derived from training a random forest (Figure S32).

1.8 Synthetic data validation

Synthetic field data was generated by forward simulation using the final parameter sets from each optimization process. The two optimization algorithms were run anew using the respectively generated synthetic data to calculate the goodness of fit statistics. The parameter sets retrieved by the validation were compared against the parameterization yielded by the optimization process.

1.9 Epidemiological outcome comparison

We conducted a small experiment to compare key epidemiological outcomes from the new parameterizations with the original model and that detail in a four malaria model comparison in Penny et al. 2016 (*11*). We simulated malaria in archetypical transmission and seasonality settings using the different parameterizations. The experiments were set up in a full-factorial fashion, considering the simulation options described in Table S1. Monitored outcomes were the incidence of uncomplicated, severe disease, hospitalizations, and indirect and direct malaria mortality over time and by age, prevalence over time and by age, the prevalence-incidence relationship, and the EIR-prevalence relationship. Simulations were conducted for a population of 10,000 individuals over 10 years.

Number of stochastic realization	Seasonality	Transmission (EIR)	Parameterization
10	Perennial	2	GA
.	Seasonal (sinusoidal)	10	GP-BO
.		25	GPSG-BO
.		50	
		100	
		200	
		300	

Table S1: Full experimental design in setting archetypes. Experiments were run at 36% probability that an infected individual receives effective care within 14 days.

1.10 Software

Consistent with previous calibration work, we used OpenMalaria version 35, an open source simulator written in C++ and further detailed in full in the supplement, wiki or in the original publications Smith 2006, Smith2012. Calibration was performed using R 3.6.0. For the machine learning processes, all algorithms were accessed through the mlr package version 2.17.0. The heteroskedastic Gaussian process utilised the hetGP package under version 1.1.2. The sensitivity analysis was conducted using the soboljansen function of the sensitivity package version 1.21.0 in R. All algorithms were adapted to the operating system (CentOS 7.5.1804) and computational resources available at the University of Basel Center for Scientific Computing, SciCORE, which uses a Slurm queueing system. The code of the calibration approach presented here is available on GitHub:

2 SUPPLEMENTARY TEXT 1: MALARIA TRANSMISSION MODEL

2.1 Main features

We test our calibration algorithm on OpenMalaria, an individual-based model of malaria dynamics. To provide context of the model's structure and the role of the fitted parameters (see supplementary text 1), we here briefly describe its main features and key equations. This description is adapted from that provided in Smith et al. 2012 (*20*) and Smith et al. 2006 (*23*). Full details of all model components can be found in *The American Journal of Tropical Medicine and Hygiene*, Volume 75, Issue 2 Supplement (2006).

OpenMalaria features discrete individual-based stochastic simulations of malaria in humans in 5-day time steps. Every infection and individual are characterized by a set of continuous state variables, namely, parasite densities, infection durations, and immune status. Key processes and relationships regarding the course of infection simulated by model include the attenuation of inoculations, acquired pre-erythrocytic immunity, acquired blood-stage immunity, morbidity (acute and severe) and mortality (malaria-specific and indirect), anemia, and the infection of vectors as a function of parasite densities in the human. Other model components include a vector model and a case management system. All individual components have previously been well documented (*20, 23*). A visual summary of the model with references to further details on each component is provided in Figure S1.

In our current recalibration only the original (base) model variant is used to test our new approach (*20*). Parameters estimated during the calibration process are highlighted and summarized in Table S2 at the end of this section. Other parameter values were drawn from the literature or were calibrated to separate data: for example, the empirical parasite density model of Maire et al. 2006 (*67*) was calibrated to malariatherapy (*68*) data and not recalibrated at the population level.

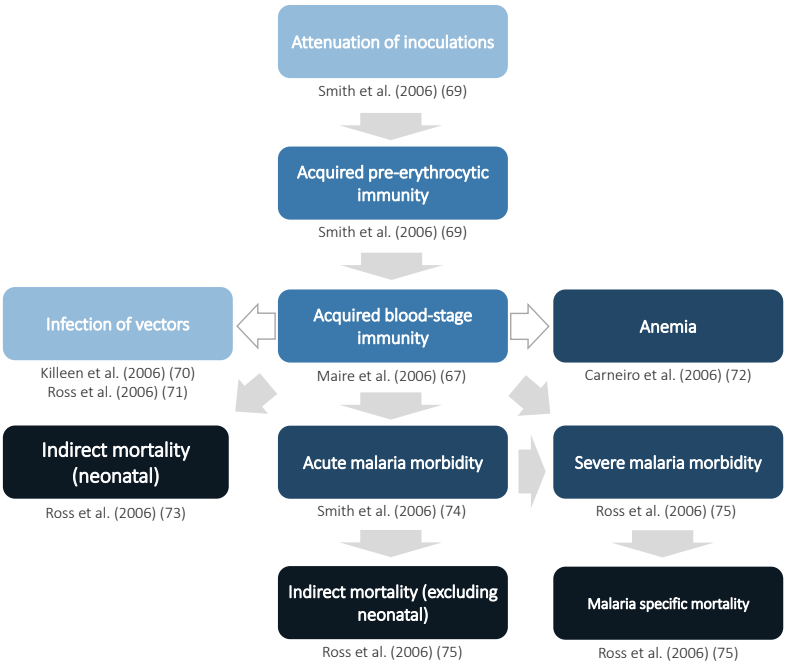


Figure S1. Visual summary of OpenMalaria with references to original publications on the model components. Adapted from Smith et al. 2006, Fig.3 (*23*). References from top to bottom and left to right (Attenuation of inoculations (*69*)),(Acquired pre-erythrocytic immunity (*69*)),(Infection of vectors [(*70*), (*71*)], Acquired blood-stage immunity (*67*), Anemia (*72*)),(Indirect mortality (neonatal) (*73*), Acute malaria morbidity (*74*), severe malaria morbidity (*75*)),(Indirect mortality excluding neonatal (*75*), Malaria specific mortality (*75*))

2.2 Infection of the human host

The seasonal pattern of entomological inoculation rate (EIR) determines seasonal pattern of transmission and thus the parasite densities in the individual, modified by natural or acquired immunity and interventions (*23*).

2.2.1 Differential feeding by mosquitoes depending on body surface area

In the base model, the expected number of entomological inoculations experienced by individual i of age a at time t is

$$E_a(i,t) = \frac{E_{max}(t)A(a(i,t))}{A_{max}} \tag{1}$$

where $E_{max}(t)$ refers to the annual entomological inoculation rate (EIR) computed from human bait collections on adults and $A()$, is the individual's availability to mosquitoes, assumed to be proportional to average body surface area, depending only on age $a(i,t)$. $A\big(a(i,t)\big)$ increases with age up to age 20 years where it reaches a value of A_{max} (the average body surface of people ≥ 20 years old in the same population).

The biting rate in relation to human weight is based on data from The Gambia published by Port and others (*76*), where the proportion of mosquitoes that had fed on a host were analyzed in relation to the host's contribution to the total biomass and surface area of people sleeping in one mosquito net (*69*).

2.2.2 Control of pre-erythrocytic stages

The number of infective bites received per unit time for each individual i, adjusted by age, is given by Eq. 1 above. A survival function $S(i,t)$ defines the probability that the progeny of an inoculation survives to give rise to a patent blood stage infection, i.e. the proportion of inoculations that result in infections or the susceptibility of individual i at time t. The force of infection is modelled as

$$\lambda(i,t) = S(i,t)E_a(i,t), \tag{2}$$

where $E_a(i,t)$ is the expected number of entomological inoculations endured by individual i at time t, adjusted for age and individual factors, and the number of infections $h(i,t)$ acquired by individual i in five-day time step t, follows a Poisson distribution:

$$h(i,t) \sim Poisson\big(\lambda(i,t)\big). \tag{3}$$

The susceptibility of individual i at time t, $S(i,t)$ is defined as:

$$S(i,t) = \left(S_\infty + \frac{1-S_\infty}{1+\frac{E_a(i,t)}{E^*}}\right)\left(S_{imm} + \frac{1-S_{imm}}{1+\left(\frac{X_p(i,t)}{X_p^*}\right)^{\gamma_p}}\right), \tag{4}$$

where $S_{imm}, X_p^*, E^*, \gamma_p$ and S_∞ are constants representing the lower limit of success probability of inoculations in immune individuals, critical value of cumulative number of entomologic inoculations, critical value of $E_a(i,t)$, steepness of relationship between success of inoculation and $X_p(i,t)$, and, the lower limit of success probability of inoculations at high where $E_a(i,t)$, respectively. Here

$$X_p(i,t) = \int_{t-a(i,t)}^{t} E_a(i,\tau)d\tau \tag{5}$$

S_∞ and E^* are fixed to $S_\infty = 0.049$, and $E^* = 0.032$ inoculations/person-night and are detailed in (*69*).

2.2.3 Course of infection in the human host

The model for each individual infection j in host i comprises a time series of parasite densities. The base model for infection within humans is described in Maire et al. 2006 (*67*). In brief, the duration of each infection, τ_{max} is sampled from

$$\ln\big(\tau_{max}(i,j)\big) \sim Normal(5.13,0.80), \tag{6}$$

parameterised against malaria therapy data (*68*) and detailed in Maire et al. 2006 (*67*). In the absence of previous exposure or concurrent infections, the log density of infection j in host i at each time point, $\tau = 0,1,\dots,\tau_{max}(i,j)$ is normally distributed with expectation

$$\ln\big(y_0(i,j,\tau)\big) = \ln d(i) + \ln y_G(\tau,\tau_{\max}), \tag{7}$$

where $y_G(\tau,\tau_{max})$ is taken from a statistical description of parasite densities in malariatherapy patients and $d(i)$ describes between-host variation with a log-normal distribution with variance σ_i^2.

We consider the possibility of multiple concurrent infections in the same individual at the same time. Exposure to asexual blood stages is measured by

$$X_y(i,j,t) = \int_{t-a}^{t} Y(i,\tau)\,d\tau - \int_{t_{0,j}}^{t} y(i,j,\tau)d\tau, \tag{8}$$

where $Y(i,\tau)$ is the total parasite density of individual i at time τ and $y(i,j,\tau)$ is the density of infection j in individual i at time τ and

$$X_h(i,t) = \int_{t-a}^{t} h(i,\tau)\,d\tau - 1. \tag{9}$$

In the presence of previous exposure and co-infection, the expected log density for each concurrent infection is then:

$$E\big(\ln\big(y(i,j,\tau)\big)\big) = D_y(i,t)D_h(i,t)D_m(i,t)\ln\big(y_0(i,j,\tau)\big) + \ln\left(\frac{D}{M(i,t)} + 1 - D_x\right), \tag{10}$$

where $M(i,t)$ is the total multiplicity of infection of in individual i at time t, and

$$D_y(i,t) = \frac{1}{1 + \frac{X_y(i,j,t)}{X_y^*}}, \tag{11}$$

where $X_y(i,j,t) = \sum_{t-a}^{t} Y(i,t) - \sum_{t_{0,j}}^{t} y(i,j,\tau)$ (note that a continuous time approximation to this is given in the original publications (*67, 69*) and hence measures the cumulative parasite load. Furthermore

$$D_h(i,t) = \frac{1}{1 + \frac{X_h(i,t)}{X_h^*}}, \tag{12}$$

where, $X_h(i,t) = \sum_{t-a}^{t} h(i,\tau) - 1$, the number of inoculations since birth, excluding the one under consideration, which measures the diversity of inocula experienced by the host up to the time point under consideration.

$$D_m(i,t) = 1 - \alpha_m \exp\left(-\frac{0.693a(i,t)}{a_m^*}\right) \tag{13}$$

which measures the effect of maternal immunity. X_y^*, X_h^*, D_x, a_m^*, and α_m are all constants estimated in the fitting process. These constants are described in Table S2, or further in Maire et al. 2006 (*67*).

Variation within individuals described as $\sigma_y^2(i,j,\tau)$, where

$$\sigma_y^2(i,j,\tau) = \frac{\sigma_0^2}{1 + \frac{X_h(i,t)}{X_v^*}}, \tag{14}$$

and σ_o^2 and X_v^* are constants, described in Table S2.

The simulated density of infection j in individual i at time τ is then drawn from a normal distribution:

$$\ln\big(y(i,j,\tau)\big) \sim Normal\ \Big(E\big(\ln\big(y(i,j,\tau)\big)\big), \sigma_y^2(i,j,\tau)\Big). \tag{15}$$

The total density of all infections in individual i at time t is then the sum of the densities of concurrent infections j

$$Y(i,t) = \sum_j y\big(i,j,\tau(i,j)\big). \tag{16}$$

2.2.4 Infectivity of the human host

The model infectivity of the human host is described in Ross 2006 where infectivity of individual I at time t is given by the distributed lag model:

$$\Upsilon(i,t) = \beta_1 Y(i,t-2) + \beta_2 Y(i,t-3) + \beta_3 Y(i,t-4), \tag{17}$$

where t is in 5-day units and

$$\ln\Big(y_g(i,t)\Big) \sim Normal\big(\ln\big(\rho\Upsilon(i,t)\big), \sigma_g^2\big), \tag{18}$$

where $\beta_1, \beta_2, \beta_3, \rho, \sigma_g^2$ are constants representing contributions of past infections to gametocyte densities (detailed in Table S2), and to be calibrated at the population level. We define

$$\Pr\left(y_g(i,t) > y_g^*\right) = \Phi\left[\frac{\ln\left(\rho\Upsilon(i,t)\right) - \ln(y_g^*)}{\sigma_g}\right] = \Phi\left[\frac{\ln\left(\Upsilon(i,t)\right)}{\sigma_g} + \rho^*\right], \tag{19}$$

where Φ is the cumulative normal distribution, y_g^* is the density of female gametocytes necessary for infection of the mosquito, and $\rho^* = \frac{\ln(\rho)-\ln(y_g^*)}{\sigma_g}$ is constant (depending on the blood meal volume, gametocyte viability and system variability). Thus, the proportion of mosquitoes infected by individual i at time t is defined as

$$I_m(i,t) = \left[\Pr\left(y_g(i,t) > y_g^*\right)\right]^2, \tag{20}$$

and the probability of a mosquito becoming infected during any feed is

$$\kappa_u(t) = \eta \frac{\sum_i A\left(a(i,t)\right) I_m(i,t)}{\sum_i A\left(a(i,t)\right)} \tag{21}$$

where η is a constant scale factor and to be calibrated.

We define $\kappa_u^{(0)}(t)$ as the value of $\kappa_u(t)$ in the simulation of an equilibrium scenario to which an intervention has been applied. Let $E_{max}^{(0)}(t+l_v)$ be the corresponding entomologic inoculation rate. $\kappa_u^{(1)}(t)$ and $E_{max}^{(1)}(t+l_v)$ are the corresponding values for the intervention scenario. Then

$$E_{max}^{(1)}(t+l_v) = \frac{E_{max}^{(0)}(t+l_v)\kappa_u^{(1)}(t)}{\kappa_u^{(0)}(t)}, \tag{22}$$

where l_v corresponds to the duration of the sporogenic cycle in the vector, which we approximate with two time steps (10 days). $\frac{E_{max}^{(0)}(t+l_v)\kappa_u^{(1)}(1)}{\kappa_u^{(0)}(t)}$ is the total vectorial capacity)

2.3 Morbidity

In order to simulate the clinical state of individual i at time t, for each five-day time step 5 independent samples from the simulated parasite density distribution are drawn for each concurrent infection j.

2.3.1 Acute morbidity (uncomplicated clinical cases)

The model for an episode of acute morbidity was originally described in (*74*) and occurs in individual i at time t with probability

$$P_m(i,t) = \frac{Y_{max}(i,t)}{Y^*(i,t) + Y_{max}(i,t)}, \tag{23}$$

where $Y^*$is the pyrogenic threshold and Y_{max} is the maximum density of five daily densities sampled during the five-day interval t.

The pyrogenic threshold changes over time following

$$\frac{dY^*(i,t)}{dt} = f_1\left(Y(i,t)\right) f_2\left(Y^*(i,t)\right) - \varpi Y^*(i,y), \tag{24}$$

where $f_1\left(Y(i,t)\right)$ is a function describing the relationship between accrual of tolerance and the parasite density $Y(i,y)$; $f_2(Y^*(i,t))$ describes the saturation of this accrual process at high values of

Y^* and $\bar{\omega}Y^*(i,t)$ determines the decay threshold with first-order kinetics, ensuring that the parasite tolerance is short-lived (*74*).

Here $f_1\big(Y(i,t)\big)$ is defined to ensure that the stimulus is not directly proportional to Y but rather that it asymptotically reaches a maximum at high values of Y:

$$f_1\big(Y(i,t)\big) = \frac{\alpha Y(i,t)}{Y_1^* + Y(i,t)}. \tag{25}$$

At high values of Y^*, a higher parasite load is required to achieve the same increase:

$$f_2\big(Y^*(i,t)\big) = \frac{1}{Y_2^* + Y^*(i,t)}. \tag{26}$$

Thus, the pyrogenic threshold $Y^*$is defined to follow

$$\frac{dY^*(i,t)}{dt} = \frac{\alpha Y(i,t)}{(Y_1^* + Y(i,t))(Y_2^* + Y^*(i,t))} - \bar{\omega}Y^*(i,t), \tag{27}$$

and the initial condition $Y^*(i,0) = Y_0^*$ at the birth of the host, where α, $\bar{\omega}$ Y_0, Y_1^* and Y_2^* are targets of the calibration, and are defined in Table S2.

2.3.2 Severe disease

The model for severe disease was described in Ross et al 2006 (*75*) and two different classes of severe episodes are considered by the model, $B_1$and B_2. $P_{B1}(i,t)$ is the probability that an acute episode (A) is of class B_1 and

$$P_{B_1}(i,t) = \Pr(H(i,t) \in B_1 | H(i,t) \in A) = \frac{Y_{max}(i,t)}{Y_{B1}^* + Y_{max}(i,t)}, \tag{28}$$

where Y_{B1}^* is a constant to be calibrated and $H(i,t)$ is the clinical status of individual i at time t.

Class B_2 of severe malaria episodes occurs when an otherwise uncomplicated episode coincides with some other insult, which occurs with risk

$$F\big(a(i,t)\big) = \frac{F_0}{1 + \left(\frac{a(i,t)}{a_F^*}\right)}, \tag{29}$$

where F_0 is the limiting value of $F\big(a(i,t)\big)$ at birth and a_F^* is the age at which it is halved and both are to be calibrated.

The probability that individual i experiences an episode belonging to class B_2 at time t, conditional on there being a clinical episode at that time is

$$P_{B_2}(i,t) = \Pr(H(i,t) \in B_2 \,|\, H(i,t) \in A) = F(a(i,t)). \tag{30}$$

The age ant time specific risk of severe malaria morbidity conditional on a clinical episode is then given by

$$P_B(i,t) = P_{B_1}(i,t) + P_{B_3}(i,t) - P_{B_1}(i,t)P_{B_2}(i,t). \tag{31}$$

2.3.3 Mortality

Malaria deaths in hospital are a random sample of admitted severe malaria cases, with age-dependent sampling fraction $Q_h(a)$, the hospital case fatality rate, derived from the data of Reyburn et al (2004) (*77*). The original model was described in Ross et al. 2006 (*75*).

The severe malaria case fatality in the community for age group a, $Q_c(a)$ is estimated as

$$Q_c(a) = \frac{Q_h(a)\phi_1}{1 - Q_h(a) + Q_h(a)\phi_1}, \tag{32}$$

where ϕ_1 the estimated odds ratio for death in the community compared to death in in-patients is an age-independent constant to be calibrated and $Q_h(a)$ is the hospital case fatality rate. The total malaria mortality is the sum of the hospital and community malaria deaths.

The risk of neonatal mortality attributable to malaria (death in class D_1) in first pregnancies is set equal to $0.3\mu_{PG}$ where

$$\mu_{PG} = \mu_{max}\left[1 - \exp\left(-\frac{x_{PG}}{x_{PG}^*}\right)\right], \tag{33}$$

where x_{PG} is related to x_{MG}, the prevalence in simulated individuals 20-24 ears of age via

$$x_{PG} = 1 - \frac{1}{1 + \left(\frac{x_{MG}}{x_{MG}^*}\right)} \tag{34}$$

and x_{MG}^* and x_{PG}^* are constants to be calibrated and are detailed in Table S2.

An indirect death in class D_2 is provoked at time t, conditional on there being a clinical episode at that time with probability $P_{D2}(i,t)$ where

$$P_{D_2}(i,t) = \Pr(H(i,t) \in D_2 | H(i,t) \in A), \tag{35}$$

and

$$P_{D_2}(i,t) = \frac{Q_D}{1 + \left(\frac{a(i,t)}{a_F^*}\right)}, \tag{36}$$

where Q_D is limiting value of $P_{D_2}(i,t)$ at birth and a_F^* is a constant to be calibrated. Deaths in class D_2 occur 30 days (six time steps) after the provoking episode.

No.*	θ^+	Parameter	Meaning	Unit/ dimension	Prior	GA-O estimate (Smith et al. 2012, model R0001)(*20*)	New estimate GP-BO (Reiker et al.2020)	New estimate GPSG-BO (Reiker et al.2020)
1	--	$-\ln(1-S_\infty)$	S_∞ = Lower limit of success probability of inoculations at high $E_a(i,t)$	Proportion	--	0.051	0.051	0.051
2	--	E^*	Critical value of $E_a(i,t)$	Inoculations/ person-night	--	0.032	0.032	0.032
3	1[a]	S_{imm}	Lower limit of success probability of inoculations in immune individuals	Proportion	$\exp(N(\log(0.14),2))$	0.138	0. 196	0.036
4	3	X_p^*	Critical value of cumulative number of entomologic inoculations	Inoculations	$\exp(N(\log(1514),2))$	1,514.4	1,954.8	4,972.2
5	2	γ_p	Steepness of relationship between success of inoculation and $X_p(i,t)$	Dimensionless constant	$\exp(N(\log(1),1))$	2.037	1.291	1.871
6	23	σ_i^2	Variation between hosts on parasite densities (variance of log-normal distribution)		$\exp(N(\log(10.17),0.6))$	10.174	11.729	9.689
7	5	X_y^*	Critical value of cumulative number of parasite days	Parasite-days/μL x 10^{-7}	$\exp(N(\log(3.52x10^7),2))$	3.516	593.661	1.216
8	4	X_h^*	Critical value of cumulative number of infections	Infections	$\exp(N(\log(97.3),2))$	97.335	54.082	89.759
9	7	$\ln(1-\alpha_m)$	α_m = Maternal protection at birth	Dimensionless	$-\log(1-Beta(8,2))$	2.330	1.770	1.266
10	8	α_m^*	Decay of maternal protection	Per year	$\exp(N(\log(1.8),0.5))$	2.531	1.279	1.551

11	9	$\boldsymbol{\sigma_0^2}$	Fixed variance component for densities	$[\mathbf{ln}(\mathbf{density})]^2$	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{0.66}),\mathbf{2})\big)$	0.656	5.838	1.440
12	6	$\boldsymbol{X_v^*}$	Critical value of cumulative number of infections for variance in parasite densities	Infections	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{5}),\mathbf{1})\big)$	0.916	3.959	7.226
13	14	$\boldsymbol{Y_2^*}$	Critical value of $\boldsymbol{Y^*(i,t)}$ in determining increase in $\boldsymbol{Y^*}$	Parasites/$\boldsymbol{\mu}$L	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{5000}),\mathbf{1})\big)$	6,502.26	6,560.08	13,485.57
14	10	$\boldsymbol{\alpha}$	Factor determining increase in $\boldsymbol{Y^*(i,t)}$	$\mathbf{Parasites^2}\boldsymbol{\mu L^{-2}}\mathbf{day^-}$	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{142602}),\mathbf{1})\big)$	142,602	63,220.5	119,502
15	22	$\boldsymbol{\nu_1}$	Density bias (non Garki)	Dimensionless	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{0.177}),\mathbf{0.6})\big)$	0.177	0.123	0.159
16	--	$\boldsymbol{\sigma_2}$	Mass action parameter	Dimensionless		1	1	1
17	18	$\mathbf{log}\,\boldsymbol{\phi_1}$	Case fatality for severe episodes in the community compared to hospital	Log odds	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{2.09}),\mathbf{0.3})\big)$	0.736	0.340	0.285
18	20[b]	$\boldsymbol{Q_D}$	Co-morbidity intercept relevant to indirect mortality	Proportion	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{0.019}),\mathbf{1})\big)$	0.019	0.019	0.023
19	19[c]	$\boldsymbol{Q_n}$	Non-malaria intercept for infant mortality	Deaths / 1000 live births	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{49.5}),\mathbf{1})\big)$	49.539	46.5095	40.163
20	21	$\boldsymbol{\nu_0}$	Density bias (Garki)	Dimensionless	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{4.79}),\mathbf{0.2})\big)$	4.796	3.739	5.618
21	15	$\boldsymbol{Y_{B_1}^*}$	Parasitaemia threshold for severe episodes type $\mathbf{B_1}$	Parasites/$\boldsymbol{\mu}$L	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{250000}),\mathbf{0.8})\big)$	784,456	849,046	484,122
22	--	--	Immune penalty		--	1	1	1
23	--	--	Immune effector decay		--	0	0	0
24	16[d]	$\boldsymbol{F_0}$	Prevalence of co-morbidity/susceptibility at birth relevant to severe episodes ($\boldsymbol{B_2}$)	proportion	$\mathbf{exp}\big(\boldsymbol{N}(\mathbf{log}(\mathbf{0.092}),\mathbf{0.5})\big)$	0.097	0.078	0.094

25	11	$\frac{\log 2}{\bar{\omega}}$	Y^*(pyrogenic threshold) half-life	Years	$\log(2)$ / $\exp(N(\log(2.52), 1))$	0.275	0.468	0.516
26	13	Y_1^*	Critical value of parasite density in determining increase in Y^*	Parasites/μL	$\exp(N(\log(6), 2))$	0. 597	1.665	0.477
27	--	--	Asexual immunity decay		--	0	0	0
28	12	Y_0^*	Pyrogenic threshold at birth	Parasites/μL	$\exp(N(\log(296.3), 1))$	296.302	90.938	201.671
29	--	--	Idete multiplier	Dimensionless	--	2.798	2.799	2.799
30	17	a_F^*	Critical age for co-morbidity	Years	$\exp(N(\log(0.225), 0.8))$	0.117	0.138	0.087

[*] Parameter number assigned for simulations in OpenMalaria scenarios, some parameters here are used in model variants and not in the base model. Listed for completeness; [+]Parameter number θ_i assigned for the optimisation problem. θ is drawn from the unit cube and determines the quantiles of the prior for the parameter value. [a] quantile = $\theta * 0.8372102$. [b] quantile = $\theta * 0.9999991$. [c] quantile = $\theta * 0.9986755$. [d] quantile = $\theta * 0.999963$.

Table S2: Names and details of OpenMalaria core parameters. GA-O = Genetic algorithm optimization, GP-BO = Gaussian process-based Bayesian optimization, GPSG-BO = Gaussian process stacked generalization-based Bayesian optimization

3 SUPPLEMENTARY TEXT 2: CALIBRATION APPROACH AND DATA SUMMARY

A comprehensive epidemiological calibration dataset was collated in order to parameterize OpenMalaria. This calibration dataset covers a total of eleven different epidemiological relationships (or objectives for fitting) that span important aspects of the natural history of malaria. Data were collated from different settings (see Table S3 for summary) and were detailed in the original model descriptions (*23, 74*) and a later parameterization (*20*). A total of 61 simulation scenarios were setup to parameterize OpenMalaria, constructed to simulate the study surveys and study sites that yielded the calibration dataset. The study site observations were replicated in OpenMalaria by reproducing the timing of the surveys and their endpoints (such as prevalence and incidence) and matching simulation options to the setting with regards to transmission intensity and seasonality, vector species, treatment seeking behavior and anti-malarial interventions. The objectives and data are further detailed below.

The parameter estimation process is a multi-objective optimization problem with each of the epidemiological quantities in Table S3 representing one objective. The aim of the optimization is to find a parameter set that maximizes the goodness of fit by minimizing a loss statistic computed as the weighted sum of the loss functions for each objective. Building a weighted average reduces the multiple loss terms to a single overall loss statistic, defined as:

$$F(\theta) = \sum_i w_i \sum_j f_{ij}(\boldsymbol{\theta}) \tag{37}$$

where $f_{ij}(\boldsymbol{\theta})$ is the loss function for parameter vector $\boldsymbol{\theta}$, epidemiological quantity i and dataset j, and the weights w_i were chosen so that different epidemiological quantities contribute approximately equally to $F(\boldsymbol{\theta})$.

For the current calibration, we utilised the loss functions from Smith et al. 2012 (*20*), the loss function $f_i(\boldsymbol{\theta})$ for each objective i use either (negative) log-likelihoods or Residual Sum of Squares (RSS) with an unknown minimum. We did not update these loss-functions in order to compare to our previous approaches.

The likelihood functions are given by

$$\mathcal{L}(\boldsymbol{\theta}|x_1, \dots, x_n) = g(x_1, \dots, x_n|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} g(x_i|\boldsymbol{\theta}) \tag{38}$$

where the observed values are $x_1, \dots, x_n$ and the model parameters $\boldsymbol{\theta}$. In practice, it is easier to work with the log likelihood, namely

$$\log \mathcal{L}(\boldsymbol{\theta}|x_1, \dots, x_n) = \sum_{i=1}^{n} \log g(x_i|\boldsymbol{\theta}) \tag{39}$$

The loss functions $f_i(\boldsymbol{\theta})$ used for each objective are detailed in the following sections.

3.1 Objectives: Epidemiological data and loss functions

Below we described each fitting objective in terms of the data (setting, surveys, observations, references) along with the associated loss function and original references. Table S4 provides an overview of the 61 simulation scenarios used for calibration, and which objective they contribute to.

3.1.1 Age pattern of incidence after intervention

3.1.1.1 Data

The data used for the calibration of objective 1 (Age pattern of incidence) consists of eight cross-sectional surveys of infection rates by age and EIR in Matsari village, capturing 12 age groups each. Matsari village was monitored entomologically for four years (Nov 1970 - Nov1973) during the Garki Project and multiple anti-malaria interventions were administered (*78*). From October 1970 to March 1972 (the baseline / pre-intervention phase), eight cross-sectional malariologic surveys of the whole village population and intensive entomologic surveillance (human bait collection of mosquitoes and dissections of the mosquito salivary glands for sporozoites) were carried out. The latter was used to estimate a baseline transmission intensity of 67 inoculations per person per year (EIR) and to derive seasonal transmission patterns. Mid-1972 marked the beginning of the intervention phase, during which an additional eight surveys were carried out at 10-week intervals (surveys 9-16). During this time, indoor residual spraying with Propoxur was carried out comprehensively in the village, along with mass treatment of the population with Sulfadoxine-pyrimethamine at 10 week-intervals immediately after assessment of individuals' parasitologic status. The experimental setup is summarised in Figure 3 of Smith et al 2006 (*69*). Incidence data (number of patent infections and number of hosts by age) from surveys 9-16 was used for our calibration.

Sites and scenario numbers: Matsari, Nigeria (30)

Original reference detailing data and model fits: Smith TA, Maire N, Dietz K, Killeen GF, Vounatsou P et al. Relationship between the entomological inoculation rate and the force of infection for Plasmodium falciparum malaria. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (69)

3.1.1.2 Loss function: Binomial Log Likelihood

We denote the Binomial log likelihood for this objective to be

$$f_1(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) = \sum_{j=1}^{s}\sum_{k=1}^{a} P_{j,l} \log(\widehat{p_j, k}) + \left(H_{j,k} - P_{j,k}\right) \log(1 - \widehat{p_{j,k}}) \tag{40}$$

where a is the number of age groups, s the number of surveys, $p_{j,k}$ the scenario data number of parasite positive hosts and $H_{j,k}$ the scenario data number of hosts for age group k and survey j. Parameter $\widehat{p_{j,k}}$ is associated with the model predictions and is given by

$$\widehat{p_{j,k}} = \widehat{P_{j,k}} \,/\, \widehat{H_{j,k}} \tag{41}$$

where $\widehat{P_{j,k}}$ are the predicted number of parasite positive hosts and $\widehat{H_{j,k}}$ the predicted number of hosts for age group k and survey j.

3.1.2 Age patterns of prevalence

3.1.2.1 Data

The data used for the calibration of objective 2 (age-patterns of prevalence) consists of six cross-sectional malariology surveys conducted in the Rafin Marke, Matsari, Sugungum villages in Nigeria

1970-1972 (12 age groups each, part of the Garki Project during the pre-intervention period) (*78*), Navrongo in Ghana 2000 (12 age groups) (*79*) and Namawala 1990-1991 (*80*) and Idete in Tanzania (11 and 6 age groups, respectively) 1992-1993 (*81*). In all study sites, annual transmission intensity (EIR) and seasonal patterns were assessed using light trap or human night bait collections and dissections of the salivary glands (see Figure 2 in Maire et al. 2006 (*67*)). In all sites except Idete, the health system at the time of the surveys treated only a small proportion of the clinical malaria episodes. In the Idete, the village dispensary was assumed to treat approximately 64% of clinical malaria (based on the published literature). During simulation, prevalence was defined by comparing each predicted parasite density with the limit of detection used in the actual study.

Sites and scenario numbers: Sugungum, Nigeria (24); Rafin-Marke, Nigeria (28); Matsari, Nigeria (29); Idete, Tanzania (31); Navrongo, Ghana (34); Namawala, Tanzania (35)

Original reference *detailing data and model fits*: *Maire N, Smith TA, Ross A, Owusu-Agyei S, Dietz K, et al. A model for natural immunity to asexual blood stages of Plasmodium falciparum malaria in endemic areas. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (67)*

3.1.2.2 Loss function: Binomial Log Likelihood

We denote the binomial log likelihood for each scenario of this objective to be

$$f_2(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) = \sum_{j=1}^{s}\sum_{k=1}^{a} P_{j,k}\log\left(p_j, k\right) + \left(H_{j,k} - P_{j,k}\right)\log(1 - p_{j,k}) \tag{42}$$

where a is the number of age groups, s the number of surveys, $P_{j,k}$ the scenario data number of parasite positive hosts and $H_{j,k}$ the scenario data number of hosts for age group k and survey j. Parameter $p_{j,k}$ is associated with the model predictions and is given by

$$p_{j,k} = \widehat{P_{j,k}} / \widehat{H_{j,k}} \tag{43}$$

where $\widehat{P_{j,k}}$ are the predicted number of parasite positive hosts and $\widehat{H_{j,k}}$ the predicted number of hosts for age group k and survey j.

3.1.3 Age patterns of parasite density

3.1.3.1 Data

The same data sources as for objective 2 (age pattern of prevalence) were used for calibration of objective 3 (age pattern of parasite density). Parasite densities in sites that were part of the Garki project (Sugungum, Rafin-Make and Matsari, Nigeria) were recorded by scanning a predetermined number of microscope fields on the thick blood film and recording how many had one or more asexual parasites visible. These were converted to numbers of parasites visible by assuming Poisson distribution for the number of parasites per field and a blood volume of 0.5 mm^3 per 200 fields. In the other studies (Idete and Namawala, Tanzania and Navrongo, Ghana), parasites were counted against leukocytes and converted to nominal parasites/microliter assuming the usual standard of 8,000 leukocytes/microliter. The biases in density estimates resulting from these different techniques were accounted for by multiplying the observed parasite densities with constant values estimated for Garki (v_0) and non-Garki (v_1) studies to rescale them to the values in malariatherapy patients (*82*).

Sites and scenario numbers: Sugungum, Nigeria (pre-intervention, 24); Rafin-Marke, Nigeria (pre-intervention, 28); Matsari, Nigeria (pre-intervention, 29); Idete, Tanzania (31); Navrongo, Ghana (34); Namawala, Tanzania (35)

Original reference detailing data and model fits: *Maire N, Smith TA, Ross A, Owusu-Agyei S, Dietz K, et al. A model for natural immunity to asexual blood stages of Plasmodium falciparum malaria in endemic areas. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006* (*67*)

3.1.3.2 Loss function: Log-normal log likelihood

For objective 3 (age pattern of parasite densities) we denote the log-Normal log likelihood for each scenario to be

$$f_3(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) = n(\log(\rho) - \log(\sigma)) - \ 0.5\mathrm{RSS}/\sigma^2 \tag{44}$$

where n is the number of observations in the data set, $\rho = \exp(-0.5\log(2\,\pi))$, a constant from the log-normal likelihood, RSS is the residual sum of squares given by

$$\mathrm{RSS} = \sum_{j=1}^{s}\sum_{k=1}^{a}\left(\frac{\widehat{Y_{j,k}}}{\widehat{P_{j,k}}} - \log(\nu) - \frac{Y_{j,k}}{P_{j,l}}\right)^2 \tag{45}$$

and σ is the standard deviation given by

$$\sigma = \sqrt{\mathrm{RSS}/\ (n-1)} \tag{46}$$

Here, ν is the appropriate density bias, which is a fitting parameter, a is the number of age groups, s is the number of surveys, $P_{j,k}$ the scenario number of parasite positive hosts, and $Y_{j,k}$ the sum of the log densities, $\widehat{P_{j,k}}$ the predicted number of parasite positive hosts and $\widehat{Y_{j,k}}$ the predicted sum of the log densities for age group k and survey j. The density bias are fitting parameters ν_0 and ν_1.

3.1.4 Age pattern of number of concurrent infections

3.1.4.1 Data

For objective 4 (age pattern of number of concurrent infections), the dataset from Navrongo, Ghana (also used in the calibration of objectives 2 and 3) was used to calibrate to the total numbers of distinct parasite infections in one individual in each age group, and at each survey. Distinct infections were detected by polymerase chain reaction-restriction fragment length polymorphism in the sampled individuals.

Sites and scenario numbers: Navrongo, Ghana (34)

Original reference *detailing data and model fits*: *Maire N, Smith TA, Ross A, Owusu-Agyei S, Dietz K, et al. A model for natural immunity to asexual blood stages of Plasmodium falciparum malaria in endemic areas. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (67)*

3.1.4.2 Loss function: Poisson Log Likelihood

Assuming that both the data and the simulations are Poisson distributed about the correct value and thereby also allowing for over-dispersion, we denote the Poisson log likelihood for each scenario to be for the objective of age pattern of number of concurrent infections to be

$$f_4(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) = \sum_{j=1}^{s}\sum_{k=1}^{a} -Pn_{j,k}\log\left(Pn_{j,k}\ /\ \lambda_{j,k}\right) + Pn_{j,k} - \lambda_{j,k} \tag{47}$$

where a is the number of age groups, s the number of surveys, $Pn_{j,k}$ the scenario data total patent infections for age group k and survey j. Parameter $\lambda_{j,k}$ is associated with the model predictions and is given by

$$\lambda_{j,k} = \frac{\widehat{Pn_{j,k}}}{\widehat{H_{j,k}}} H_{j,k} \tag{48}$$

where $\widehat{Pn_{j,k}}$ are the predicted total of patent infections and $\widehat{H_{j,k}}$ the predicted number of hosts for age group k and survey j and $H_{j,k}$ is the scenario data number of hosts for age group k and survey j.

3.1.5 Age pattern of incidence of clinical malaria

3.1.5.1 Data

Two distinct datasets representing three study sites (Table S5) were used for the calibration of objective 5 and objective 6 (age pattern of incidence of clinical malaria). For Objective 5, the dataset contains data on the age pattern of clinical episodes in the villages of Ndiop and Dielmo in Senegal (*83, 84*). During the study period of July 1990 - June 1992, the village populations were visited daily to detect and treat any clinical malaria attacks with quinine. Cases were detected by reporting of symptoms (fever) during daily active case detection and subsequent thick blood smear microscopy. Only symptomatic individuals (axillary temperature ≥ 38.0°C or rectal temperature ≥ 38.5°C). Due to the active case detection and rapid treatment all symptomatic episodes are assumed to be effectively treated in these villages during the study period. No effective treatment of clinical malaria was assumed prior to the study period. The annual patterns of transmission were replicated as reported by Charlwood et al (1998) (*85*). A proportion P_t=35.75% are assumed to be treated effectively in Idete. As all individuals reporting to the village dispensary were treated presumptively with chloroquine, this proportion corresponds to the proportion of episodes reported to the village dispensary.

Sites and scenario numbers: Ndiop, Senegal (232), Dielmo, Senegal (233)

Original reference detailing data and model fits: *Smith TA, Ross A, Maire N, Rogier C, Trape J-F et al. An epidemiologic model of the incidence of acute illness in Plasmodium falciparum malaria. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (74)*

3.1.5.2 Loss function: RSS-biased

We denote a loss function based on biased residual sum of squares:

$$f_5(\boldsymbol{\theta}) = \sum_{j=s_1}^{s} \sum_{k=1}^{a} R^2 \tag{49}$$

where a is the number of age groups, s the number of surveys, s_1 the initial survey number, and R is the residual given by

$$R = I_{i,j} - \frac{\widehat{C_{j,k}}}{\left(\widehat{H_{j,k}}\right)} \frac{1}{\mu} \tag{50}$$

where $I_{j,k}$ is the observed recorded incidence rate, $\widehat{C_{j,k}}$ are the predicted total cases (severe and uncomplicated), $\widehat{H_{j,k}}$ the predicted number of hosts for age group k and survey j and μ is a bias related to the scenario. For scenarios 232 and 233 (representing Ndiop and Dielmo, Senegal) this bias is $\mu = 5$ indicating the duration in years for which episodes are collected. For scenario 49 in Objective 6 (Idete, Tanzania) the bias is $\mu = 0.357459$ and represents the proportion of episodes reported to the village dispensary.

Scenario No.	Study site	Age groups	Observations
232	Ndiop, Senegal	22	One per age group
233	Dielmo, Senegal	22	One per age group
49	Idete, Tanzania	4	One per age group

Table S5: summary of study data set for objective 5: Age pattern of incidence of clinical malaria.

3.1.6 Age pattern of incidence of clinical malaria: infants

3.1.6.1 Data

Objective 6 (age pattern of incidence of clinical malaria in infants) is informed by a dataset on incidence that contains passive case detection data on the age-incidence in infants recorded at the health centre in Idete, Tanzania from June 1993-October 1994 (*81*). The annual patterns of transmission were replicated as reported by Charlwood et al (1998) (*85*).

Sites and scenario numbers: Idete, Tanzania (49))

Original reference: *Smith TA, Ross A, Maire N, Rogier C, Trape J-F et al. An epidemiologic model of the incidence of acute illness in Plasmodium falciparum malaria. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (74)*

3.1.6.2 Loss function: RSS-biased

The loss function for Objective 6 is the same as Objective 5. For scenario 49 (Idete, Tanzania) the bias is $\mu = 0.357459$ and represents the proportion of episodes reported to the village dispensary.

3.1.7 Age pattern of threshold parasite density for clinical attacks

3.1.7.1 Data

Objective 7 (Age pattern of threshold parasite density for clinical attacks), uses the dataset from Dielmo, Senegal (see objective 5) for calibration. The pyrogenic threshold in the (OpenMalaria) predictions is output as the sum of the log threshold values across age groups. The pyrogenic threshold per age group is given as the parasite:leucocyte ratio for recorded incidence of disease. To adjust these densities to the same scale as that used in fitting the simulation model to other datasets, the parasite:leukocyte ratios were multiplied by a factor of 1,416 to give a notional density in parasites/microliter of blood. This number was derived as follows: Parasites were counted against leukocytes and converted to nominal parasites/microliter assuming the usual (though biased) standard of 8,000 leukocytes/microliter. The biases in density estimates resulting from these different techniques was accounted for by multiplying the observed parasite densities with constant values estimated for Garki (ν_0) and non-Garki (ν_1) studies to rescale them to the values in malariatherapy patients (*82*).The value 1416 comes from

$$8000\nu_1 \tag{51}$$

where the original $\nu_1 \approx 0.18$.

Sites and scenario numbers: Dielmo, Senegal (234)

Original reference detailing data and model fits: *Smith TA, Ross A, Maire N, Rogier C, Trape J-F et al. An epidemiologic model of the incidence of acute illness in Plasmodium falciparum malaria. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (74)*

3.1.7.2 Loss function: RSS-biased (log)

For the objective 7 (Age pattern of threshold parasite density for clinical attacks) we denote a residual sum of squares loss function given by (13) with

$$f_7(\boldsymbol{\theta}) = \log\left(Y_{j,k}^*\right) - \frac{\widehat{Y_{J,k}^*}}{\widehat{H_{J,k}}} - \log(\mu) \tag{52}$$

where Y^* is the observed pyrogenic threshold, $\widehat{Y^*}$ are the predicted sum log pyrogenic threshold, $\widehat{H_{J,k}}$ the predicted number of hosts for age group k and survey j and is a bias related to the scenario. Here, this bias is related to the log parasite/leucocyte ratio and thus $\mu = 1/(8000\nu_1)$ where ν_1 is the non-Garki density bias.

3.1.8 Hospitalization rate in relation to prevalence in children

3.1.8.1 Data

Data on the relative incidence of severe malaria-related morbidity and mortality in children <9 years old across different transmission intensities were originally collated by Marsh and Snow (1999) (*86*) (Table 4). Data measurements per age group were available as the relative risk of severe disease compared to age group 1 and the proportion/prevalence of severe episodes. A total of 26 entries on the relationship between severe malaria hospital admission rates and *P. falciparum* prevalence were used to calibrate objective 8 (Hospitalisation rate in relation to prevalence in children), each represented in a separate simulation scenario, with one observation per scenario. These are summarised in Table S6. To obtain a continuous function relating hospital incidence rates to prevalence, linear interpolation between data points was performed. To convert hospital incidence rates to community severe malaria incidence, the hospital admission rates was divided by the assumed proportion of severe episodes representing to hospital (48%). There was assumed to be no effective treatment of uncomplicated malaria episodes or malaria mortality.

Sites and scenario numbers: Bo, Sierra Leone (501); Niakhar, Senegal (502), Farafenni, The Gambia (503); Areas I-V, The Gambia (504-508); Gihanga, Burundi (509); Katumba, Burundi (510); Karangasso, Burkina Faso (511); Kilifi North, Kenya (512); Manhica, Mozambique (514); Namawala, Tanzania (515); Navrongo, Ghana (516); Saradidi, Kenya (517); Yombo, Tanzania (518); Ziniare, Burkina Faso (519); Matsari, Nigeria (520); ITC control, Burkina Faso (521); Mlomp, Senegal (522); Ganvie, Benin (523); Kilifi Town, Kenya (524); Chonyi, Kenya (525); Bandafassi, Senegal (526); Kongodjan, Burkina Faso (527)

Original reference detailing data and model fits: *Ross A, Maire N, Molineaux L and Smith TA. An epidemiologic model of severe morbidity and mortality caused by Plasmodium falciparum. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (75)*

3.1.8.2 Loss function: squared deviation

The loss function is denoted as the log of residual sum of squares

$$f_8(\boldsymbol{\theta}) = \left[\log\left(\frac{a_s\widehat{R_{k=1}}}{R_{k=1}^*}\right)\right]^2 \tag{53}$$

where a_s is the access to treatment of severe cases (0.48, estimated in base model), $\widehat{R_{k=1}}$ is the scenario predicted rate of severe episodes per 1000 person-years for age group $k = 1$ (0-9 years), and parameter $R^*_{k=1}$ is the interpolated observed rate of severe episodes per 1000 person year given by

$$R^*_{k=1} = \frac{\left(\widehat{P_{k=1}} - P_l\right)}{(P_u - P_l)}(R_u - R_l) + R_t \tag{54}$$

where $\widehat{P_{k=1}}$ is the predicted prevalence summed over all surveys, P_u and P_l are the observed prevalences above and below the predicted prevalence $\widehat{P_{k=1}}$, respectively and R_u and $R_l$are the corresponding severe episode rates to the observed prevalences.

The predicted prevalence is given by

$$\widehat{P_{k=1}} = \frac{\widehat{Pt_{k=1}}/24}{\widehat{H_{k=1}}/24} \tag{55}$$

where $\widehat{P_{k=1}}$ is the total number of parasite positive predicted and $\widehat{H_{k=1}}$ are the total number of hosts (division by 24 to give mean values). The predicted rate of episodes per 1000 person year is given by

$$\widehat{R_{k=1}} = \frac{1000\ \widehat{S_{k=1}}/2}{\widehat{H_{k=1}}/24} \tag{56}$$

where $\widehat{S_{k=1}}$ is the number of severe cases predicted and with division by 2 to convert to from 2 years to 1 year and the division by 24 to give mean number of hosts.

Site	**EIR data**	
	Year	**EIR**
Burkina Faso		
ITC Control	1994-1995	389
Karangasso	1985	244
Kongodjan	1984	133
Ziniare	1994-1995	70
Burundi		
Gihanga	1983	205
Katumba	1982	13.6
Kenya		
Chonyi	1992-1993	50
Kilifi North	1992-1003	10.5
Kilifi Town	1990-1991	2.8
Saradidi	1986-1987	239
Senegal		

Bandafassi	1995-1996	363
Mlomp	1995	30
Niakhar	1995	11.6
Tanzania		
Namawala	1990-1991	329
Yombo	1992	234
The Gambia		
Area I-V	1991	+
Farafenni	1987	8.9
Others		
Bo, Sierra Leone	1990-1991	34.7
Ganvie, Benin	1993-1995	11
Manhica, Mozambique	2001-2002	38
Matsari, Nigeria	1971	68
Navrongo, Ghana	2001-2002	418

*EIR = entomological inoculation rate, ITC = control group of randomised trial of insecticide-treated curtains. +Five sites with annual EIR between 1 and 10

Table S6. Settings used for calibrating the incidence of severe malaria. (Adapte from Table1 from Ross et al. 2006 (*75*))

3.1.9 Age pattern of hospitalization: severe malaria

3.1.9.1 Data

For objective 9 (Age pattern of hospitalisation), a subset of the data collated by Marsh and Snow (1999) (*86*) (see objective 8) is used. Detailed age-specific severe hospital admission rates were available for 5 of the sites (Table S7). The patterns of incidence by age were summarised by age in 1-4 and 5-9 year-old children and compared with 1-11 month old infants by calculating the relative risk. Of the five sites, four were selected for fitting objective 9 based on the predicted prevalence. Baku, The Gambia was excluded as the very low (2%) prevalence here could not be matched.

Sites and scenario number(s): Area V, The Gambia (158); Saradidi, Kenya (167); Ganvie, Benin (173); Bandafassi, Senegal (176)

Original reference detailing data and model fits: *Ross A, Maire N, Molineaux L and Smith TA. An epidemiologic model of severe morbidity and mortality caused by Plasmodium falciparum. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (75)*

Estimate	Sukuta, The Gambia	Kilifi North, Kenya	Kilifi South, Kenya	Siaya, Kenya
Years of paediatric ward surveillance	1992-95	1990-95	1992-96	1992,1994-96

Person-years exposure to risk of children aged 0-9 yr	23468	52675	45967	40064
Rates				
All-cause malaria, age 1-11 mo	23.3 (17.8–28.9) [66/2830]	59.5 (53.2-65.9) [318/5342]	79.9 (71.6-86.4) [407/5152]	84.6 (76.4-92.8) [374/4420]
All-cause malaria, age 1-4 yr	35.3 (32.2-39.4) [372/10379]	41.7 (39.0-44.4) [905/21714]	17.4 (15.5-19.3) [321/18493]	18.8 (16.7-20.9) [312/16567]
All-cause malaria, age 5-9 yr	16.3 (13.8-18.8) [167/10259]	5.3 (4.4-6.2) [135 / 25619]	1.7 (1.2-2.2) [38/22322]	1.7 (1.1-2.3) [33/19077]
All-cause malaria, age 0-9 yr	25.8 (23.8-27.8) [605]	25.9 (24.5-27.2) (1363) [+]	16.7 (15.5-17.9) [766]	18.0 (16.7-19.3) [719]
Cerebral malaria 0-9 yr	2.6 (2.0-3.3) [61]	1.5 (1.2-1.8) [79]	0.8 (0.5-1.1) [36]	0.1 (0.0-0.2) [5]
Severe malaria anaemia, 0-9 yr	NA	5.0 (4.4-5.6) [262]	4.2 (3.6-4.8) [192]	3.7 (2.7-4.7) [50/13416]
All-cause ARI age 0-9 yr	8.4 (7.3-9.6) [198]	9.3 (8.5-10.1) [492]	8.3 (7.5-9.1) [380]	8.7 (7.8-9.6) [348]

* Period prevalence rather than incidence because precise matching of each community member to hospital admission was not possible. Rates as admission per 1000 children per year (95% CI). [+]Precise dates of birth unobtainable for five children. Defined as child admitted with primary diagnosis of malaria and Blantyre coma score of 2 or less. Defined in child with primary diagnosis of malaria and haemoglobin of 5.0g/dL or less on admission. Rates for Siaya derived from person-years exposure to risk and admissions for period Nov 1, 1994 to Oct 31, 1995

Table S7: Age-specific period prevalence rates* of severe malaria, severe malaria, severe malaria anaemia and acute respiratory-tract infections from five communities in The Gambia and Kenya. (Adapted from Table 2 from Snow et al 1997 (*87*))

3.1.9.2 Loss function: Residual sums of squares for relative risk

We denote a loss function based on residual sum of squares:

$$f_9(\boldsymbol{\theta}) = \sum_{k=2,3} \left[\log \frac{\widehat{RR}_k}{RR_k} \right]^2 \tag{57}$$

where RR_k is the relative risk of severe episode for age group k compared to age group 1 and $\widehat{RR}_k$ is the predictive relative risk for age group k compared to age group 1. The predicted relative risk is given by

$$\widehat{RR}_k = \frac{\widehat{S_k}}{\widehat{H_k}} - \frac{\widehat{S_1}}{\widehat{H_1}} \tag{58}$$

where $\widehat{S_k}$ is the number of severe cases predicted for age group k and $\widehat{H_k}$ the total number of hosts for age group k.

3.1.10 Malaria specific mortality in children (< 5 years old)

3.1.10.1 Data

For objective 10 (Malaria specific mortality in children (<5 years old)), a subset of the data collated by Marsh and Snow (1999) (*86*) (see objective 8) was used (*88*). Mortality data were derived from verbal autopsy studies in sites with prospective demographic surveillance and were adjusted for the effect of malaria transmission intensity on the sensitivity and specificity of the cause of death determination. The odds ratio for death of a case in the community relative to that in hospital was estimated by fitting to the malaria-specific mortality rates in children less than five years of age assuming the published hospital case fatality rate. Nine sites for which both malaria-specific mortality rates and seasonal transmission patterns were available were included for calibration.

There is one observation per study site and simulation scenario, and predicted values are for one survey at the end of 2 years.

Sites and scenario number(s): Bo, Sierra Leone (301); Niakhar, Senegal (302); Farafenni, The Gambia (303); Kilifi North, Kenya (312); Navrongo, Ghana (316); Saradidi, Kenya (317); Yombo, Tanzania (318); Bandafassi, Senegal (326); Kongodjan, Burkina Faso (327)

Original reference detailing data and model fits: *Ross A, Maire N, Molineaux L and Smith TA. An epidemiologic model of severe morbidity and mortality caused by Plasmodium falciparum. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006* (*75*)

3.1.10.2 Loss function: Residual sums of squares

For objective 10 on Malaria specific mortality in children, the loss function minimizes the log sum of squares

$$f_{10}(\boldsymbol{\theta}) = \left[\log\left(\frac{\widehat{DMR_1}}{DMR_1}\right)\right]^2 \tag{59}$$

where DMR_1 is the observed direct mortality rate for age group 1 (0-5 years) and $\widehat{DMR_1}$ is the predicted direct mortality rate for age group 1. The predicted direct mortality rate is given by

$$\widehat{DMR_1} = \frac{\widehat{DD_1}}{2\widehat{H_1}} \tag{60}$$

where $\widehat{DD_1}$ is the number of direct malaria deaths cases predicted for age group 1 and $\widehat{H_1}$ the total number of predicted hosts for age group 1. The division by 2 is to convert to yearly rate as the survey was conducted at the end of 2 years.

3.1.11 Indirect malaria infant mortality rate

3.1.11.1 Data

For objective 11 (indirect malaria infant mortality rate), a subset of the data collated by Marsh and Snow (1999) (*86*) (see objective 8) was used. These constitute a library of sites for which entomologic data were collected at least monthly and all-cause infant mortality rates (IMR) were available. There is one observation per scenario: all cause infant mortality rate (returned as a single number over whole intervention period).

Sites and scenario number(s): Bo, Sierra Leone (401); Niakhar, Senegal (402); Area V, The Gambia (408); Karangasso, Burkina Faso (411); Manhica, Mozambique (414); Namawala, Tanzania (415);

Navrongo, Ghana (416); Saradidi, Kanya (417); Yombo, Tanzania (418); Mlomp, Senegal (422); Bandafassi, Senegal (426)

Original reference detailing data and model fits: *Ross A, Maire N, Molineaux L and Smith TA. An epidemiologic model of severe morbidity and mortality caused by Plasmodium falciparum. Am J Trop Med Hyg. Volume 75, No. 2 Supplement. 2006 (75)*

3.1.11.2 Loss function: Residual sums of squares

The loss function minimises the log sum of squares:

$$f_{11}(\boldsymbol{\theta}) = \left[\log\left(\frac{i\widehat{DMR}_1}{iDMR_1}\right)\right]^2 \tag{61}$$

where $iDMR_1$ the observed indirect mortality rate for age group 1 and $i\widehat{DMR}_1$ is the predicted indirect mortality rate for age group 1.

3.2 Tables S3-S4

Epidemiological quantity	Data sources	No. of scenarios	No. of data points*	Publication for fitting of base model	Prior	Weighting in GOF statistic	Scenario numbers	Loss vector number (f_i)	Loss function
Age pattern of incidence of infection after intervention	Molineaux and Gramiccia (1980) (*78*)	1	12	Maire et al 2006 (*67*)	Binomial	0.001	30	1	Binomial log-likelihood
Age patterns of prevalence of infection	Molineaux and Gramiccia (1980) (*78*)	6	563	Maire et al 2006 (*67*)	Binomial	0.001	24, 28, 29, 35, 34, 31	2	Binomial log-likelihood
Age patterns of parasite density	Molineaux and Gramiccia (1980) (*78*)	6	563	Maire et al 2006 (*67*)	Log Normal	0.01	24, 28, 29, 35, 34, 31	3	log likelihood
Age pattern of number of concurrent infections	Maire et al. 2006 (*67*); Owusu-Agyei et al 2002 (*79*)	1	12	Maire et al 2006 (*67*)	Poisson	0.01	34	4	Poisson log-likelihood
Age pattern of incidence of clinical malaria: age-specific Ndiop & Dielmo, Senegal	Trape and Rogier 1996 (*83*); Kitua et al 1996 (*81*)	2	26	Smith et al 2006 (*69*)	Log Normal	1	232, 233, 49	5	RSS
Age pattern of incidence of clinical malaria: infants Idete, Tanzania	Kitua et al 1996 (*81*)	1	4	Smith et al 2006 (*69*)	Log Normal	1	49	6	RSS
Age pattern of threshold parasite density for clinical attacks	Rogier et al 1996 (*89*)	1	13	Smith et al 2006 (*69*)	Log Normal	1	234	7	RSS
Hospitalisation rate in relation to prevalence in children	See Ross et al 2006 (*75*)	26	10	Ross et al 2006 (*75*)	Log Normal	2	501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527	8	Squared deviation
Age pattern of hospitalisation: severe malaria	Marsh and Snow 1999 (*86*)	4	12	Ross et al 2006 (*75*)	Log Normal	2	158, 167, 173, 176	9	RSS

Malaria specific mortality in children (<5y)	Snow et al 1997 (*87*)	9	9	Ross et al 2006 (*75*)	Log Normal	1	301, 302, 303, 312, 316, 317, 318, 326, 327	10	Squared deviation logRate
All-cause infant mortality rate	See Ross et al 2006 (*75*)	11	11	Ross et al 2006 (*75*)	Log Normal	10	401, 402, 408, 411, 414, 415, 416, 417, 418, 422, 426	11	Squared deviation logRate

Table S3: Epidemiological quantities and data sources used for parameterizing models. (a) Some scenarios are used to predict several outcomes, so the total of this column does not equal the total of 61 scenarios involved in fitting the models. (b) The number of data points is the sum over all scenarios and simulated survey periods of the number of age groups into which the data were disaggregated for comparison with the model predictions. (c) In relation to the EIR specified as a seasonal pattern. (d) Model predictions for this objective are compared with linear interpolations between the field data points. *The number of data points is the sum over all scenarios and simulated survey periods of the number of age groups into which the data were disaggregated for comparison with the model predictions. Table adapted from Table S1 in Smith et al 2012 (*20*).

Scen. No.	Site/reference	Description	Objective(s)	Data Reference
24	Sugungum, Nigeria (pre-intervention phase)	8 cross sectional surveys of entire village population at 10-week intervals (4,487 blood slides)	Age-prevalence (2); Age-parasite densities (3)	Molineaux and Gramiccia. 1980 (*78*)
28	Rafin-Marke, Nigeria (pre-intervention phase)	8 cross sectional surveys of entire village population at 10-week intervals (2,593 blood slides)	Age-prevalence (2); Age-parasite densities (3)	Molineaux and Gramiccia. 1980 (*78*)
29	Matsari, Nigeria (pre-intervention phase)	8 cross sectional surveys of entire village population at 10-week intervals (2,963 blood slides)	Age-prevalence (2); Age-parasite densities (3)	Molineaux and Gramiccia. 1980 (*78*)
30	Matsari, Nigeria (intervention phase)	8 cross sectional surveys of entire village population at 10-week intervals (2,663 blood slides)	Age-incidence of patent infections (1)	Molineaux and Gramiccia. 1980 (*78*)
31	Idete, Tanzania	Surveillance of a rolling cohort of infants (1,382 blood slides over 16 months). Also 1 cross-sectional survey of 312 children 1-5 months	Age-prevalence (2); Age-parasite densities (3)	Kitua et al 1996 (*81*)
34	Navrongo, Ghana	6 age-stratified cross-sectional surveys at 2-month intervals (total 522 slides / DNA samples)	Age-prevalence (2); Age-parasite densities (3),	Owusu-Agyei S et al. 2002 (*79*)

			Age-specific multiplicity of infection (4)	
35	Namawala, Tanzania	12 age-stratified cross-sectional surveys at 2-month intervals (3,901 blood slides)	Age-prevalence (2); Age-parasite densities (3)	Smith et al 1993 (*80*)
49	Idete, Tanzania	Passive case detection at the village dispensary over 15 months in 12 age groups.	Age Pattern of Incidence of Clinical Malaria in Idete in infants (5b)	Kitua et al. 1996 (*81*); Vounatsou et al. 2000 (*90*)
158	Area V, The Gambia	Hospitalisation rate by age	Age pattern of severe hospitalisation (8)	Snow et al. 1997 (*87*)
167	Saradidi, Kenya	21 cohorts each of approximately 50 children between 6 months and 6 years of age whose parasites were cleared and who were then followed up with 2 weekly surveys. Hospitalisation rate by age.	Age pattern of severe hospitalisation (8)	Beier et al. 1999 (*91*), Snow 1997 (*87*)
173	Ganvie, Benin	Hospitalisation rate by age.	Age pattern of severe hospitalisation (8)	Snow et al. 1997 (*87*)
176	Bandafassi, Senegal	Hospitalisation rate by age.	Age pattern of severe hospitalisation (8)	Snow et al. 1997 (*87*)
232	Ndiop, Senegal	Longitudinal study of 350 permanent residents over 2 years: Individual level active case detection three times a week (questionnaire + recording of symptoms) and parasitologic surveys twice a week; daily recording of new fever cases at compound level. By age group (9 groups)	Age pattern of incidence of clinical malaria (5a)	Trape JF and Rogier C. 1996 (*83*)
233	Dielmo, Senegal	Longitudinal study of 206 permanent residents over 2 years: Individual level active case detection three times a week (questionnaire + recording of symptoms) and parasitologic	Age pattern of incidence of clinical malaria by age (5a)	Trape JF and Rogier C. 1996 (*83*)

		surveys twice a week; daily recording of new fever cases at compound level. By age group (9 groups)		
234	Dielmo, Senegal	Longitudinal study of 206 permanent residents over 2 years: Individual level active case detection three times a week (questionnaire + recording of symptoms) and parasitologic surveys twice a week; daily recording of new fever cases at compound level. By age group (9 groups)	Age Pattern of parasite density threshold for clinical attack (6)	Trape JF and Rogier C. 1996 (*83*)
301	Bo, Sierra Leone	Point estimate based on a 1-year longitudinal study covering 776 person-years	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
302	Niakhar, Senegal	Point estimate based on 5-year longitudinal study covering 29,491 person-years [XML label: Diohine]	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
303	Farafenni, The Gambia	Point estimate based on 2-year longitudinal study covering 2,263 person-years [XML label: Tally Ya]	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
312	Kilifi North, Kenya	Point estimate based on 3-year longitudinal study covering 20,679 person-years	Direct Malaria Mortality (9)	Korenromp et al. 2003(*88*)
316	Navrongo, Ghana	Point estimate based on 1-year longitudinal study covering 1,065 person-years	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
317	Saradidi, Kenya	21 cohorts each of approximately 50 children between 6 months and 6 years of age whose parasites were cleared and who were then followed up with 2 weekly surveys.	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
318	Yombo, Tanzania	Point estimate based on 3-year longitudinal study covering 5,850 person-years	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
326	Bandafassi, Senegal	Point estimate based on 6-year longitudinal study covering 8,488 person-years	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)

327	Kongodjan, Burkina Faso	Point estimate based on 5-year longitudinal study covering 1,271 person-years	Direct Malaria Mortality (9)	Korenromp et al. 2003 (*88*)
401	Bo, Sierra Leone	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	Barnish et al. 1993 (*92*)
402	Niakhar, Senegal	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates; XML label: Diohine	All-cause mortality (10)	INDEPTH Network, 2002 (*93*); Spencer et al. 1987 (*94*)
408	Area V, The Gambia	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	D'Alessandro et al. 1995 (*95*)
411	Karangasso, Burkina Faso	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	Duboz et al. 1989 (*96*)
414	Manhica, Mozambique	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	INDEPTH Network, 2002 (*93*)
415	Namawala, Tanzania	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates; Pre-intervention	All-cause mortality (10)	Armstrong-Schellenberg et al. 1999 (*97*)
416	Navrongo, Ghana	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	INDEPTH Network, 2002 (*93*)
417	Saradidi, Kanya	21 cohorts each of approximately 50 children between 6 months and 6 years of age whose parasites were cleared and who were then followed up with 2 weekly surveys.	All-cause mortality (10)	Spencer et al. 1987(*94*)
418	Yombo, Tanzania	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	Premji Z et al. 1997 (*98*)
422	Mlomp, Senegal	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	Trape et al. 1998 (*99*)
426	Bandafassi, Senegal	Point estimates of all-cause neonatal, post-neonatal, and infant mortality rates	All-cause mortality (10)	INDEPTH Network, 2002 (*93*)

501	Bo, Sierra Leone	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
502	Niakhar, Senegal	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.XML label: Diohine (ca 20 km from Niakhar)	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
503	Farafenni, The Gambia	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.XML label: Tally Ya (ca 15 km from Farafenni)	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
504	Area I, The Gambia	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
505	Area II, The Gambia	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
506	Area III, The Gambia	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
507	Area IV, The Gambia	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
508	Area V, The Gambia	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
509	Gihanga, Burundi	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
510	Katumba, Burundi	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
511	Karangasso, Burkina Faso	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)

512	Kilifi North, Kenya	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
514	Manhica, Mozambique	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
515	Namawala, Tanzania	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old. Pre-intervention	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
516	Navrongo, Ghana	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
517	Saradidi, Kenya	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
518	Yombo, Tanzania	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
519	Ziniare, Burkina Faso	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
520	Matsari, Nigeria	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old. Pre-intervention	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
521	ITC control, Burkina Faso	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
522	Mlomp, Senegal	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
523	Ganvie, Benin	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)

524	Kilifi Town, Kenya	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
525	Chonyi, Kenya	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
526	Bandafassi, Senegal	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)
527	Kongodjan, Burkina Faso	Point estimate of the severe malaria hospital admission rate and P.falciparum prevalence in children <9 years old.	Severe episodes by prevalence (7)	Marsh and Snow 1999 (*86*)

Table S4: Calibration data for objectives 2-4, age patterns of prevalence, parasite densities, and multiplicity of infection

4 EMULATOR PERFORMANCE

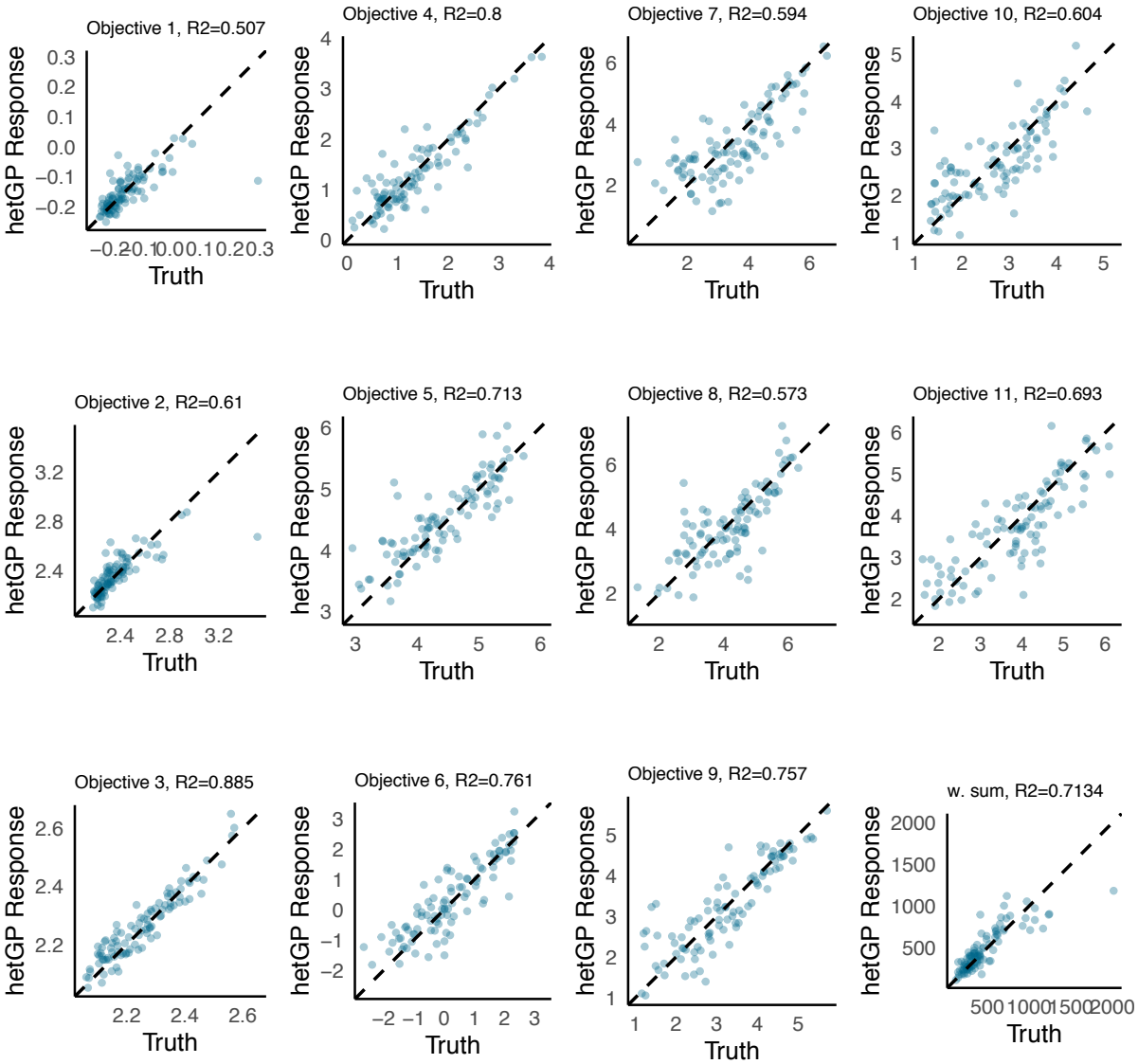


Figure S2. GP emulator performance. Emulator predictions vs true values on a holdout set compromising 10% of initial samples in iteration 1. w.sum is the weighted sum F, of the 11 objectives. Here, predictions are generated as the weighted sum of individual objective predictions.

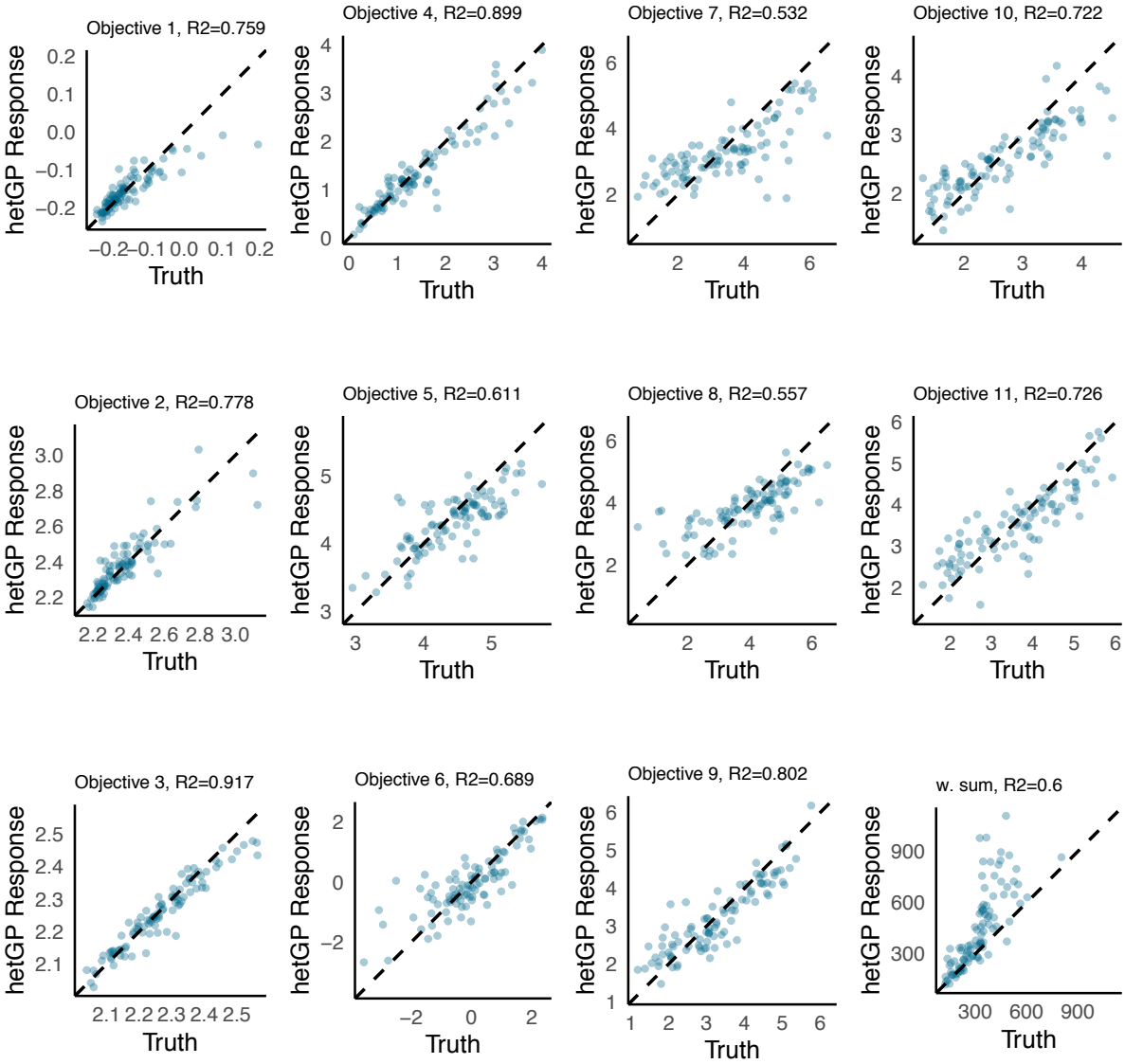


Figure S3. GP emulator performance. Emulator predictions vs true values on a holdout set compromising 10% of initial samples in iteration 30 (final iteration). w.sum is the weighted sum F, of the 11 objectives. Here, predictions are generated as the weighted sum of individual objective predictions.

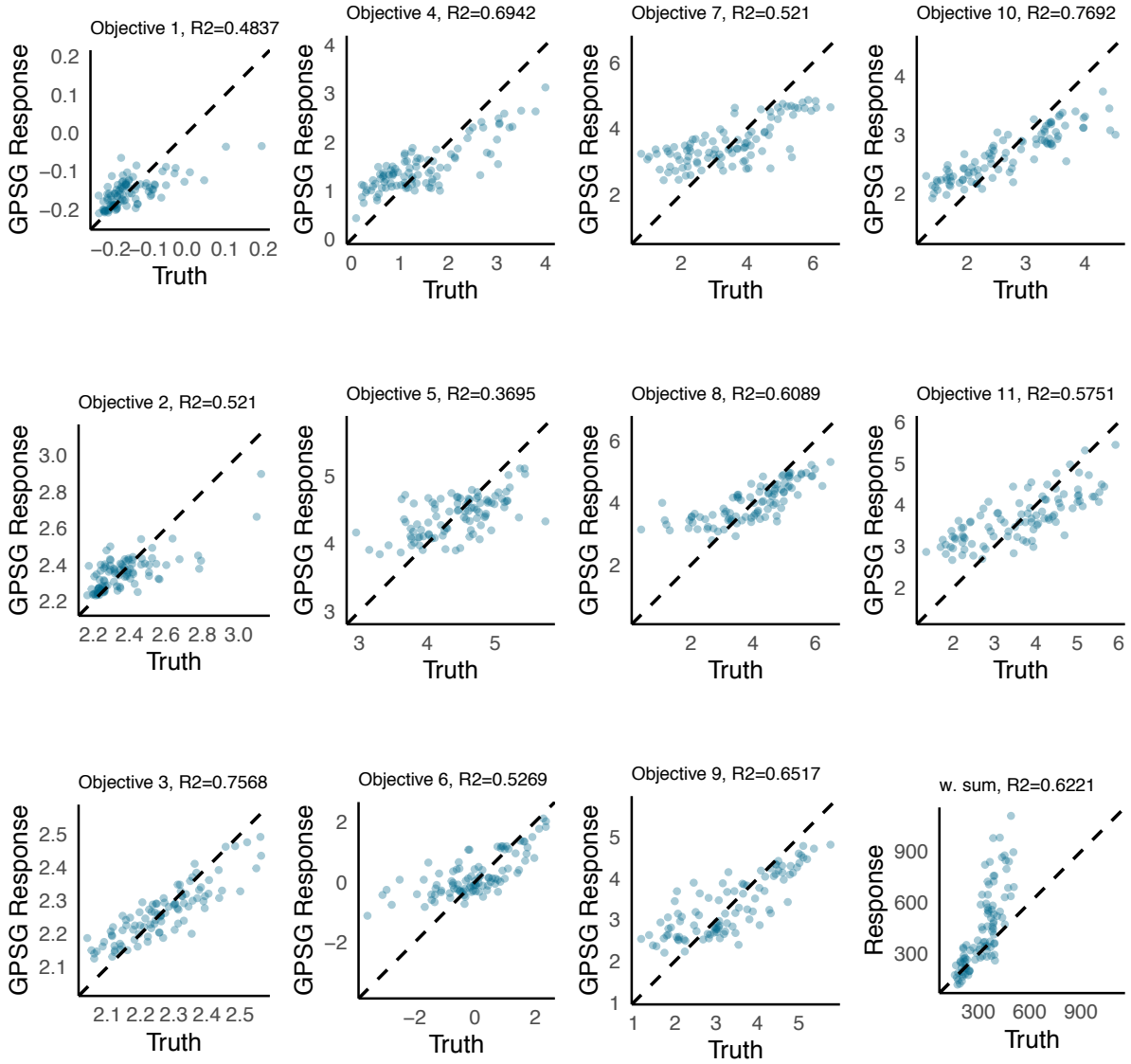


Figure S4. GPSG emulator performance. Emulator predictions vs true values on a holdout set compromising 10% of initial samples in iteration 1. w.sum is the weighted sum F, of the 11 objectives. Here, predictions are generated as the weighted sum of individual objective predictions.

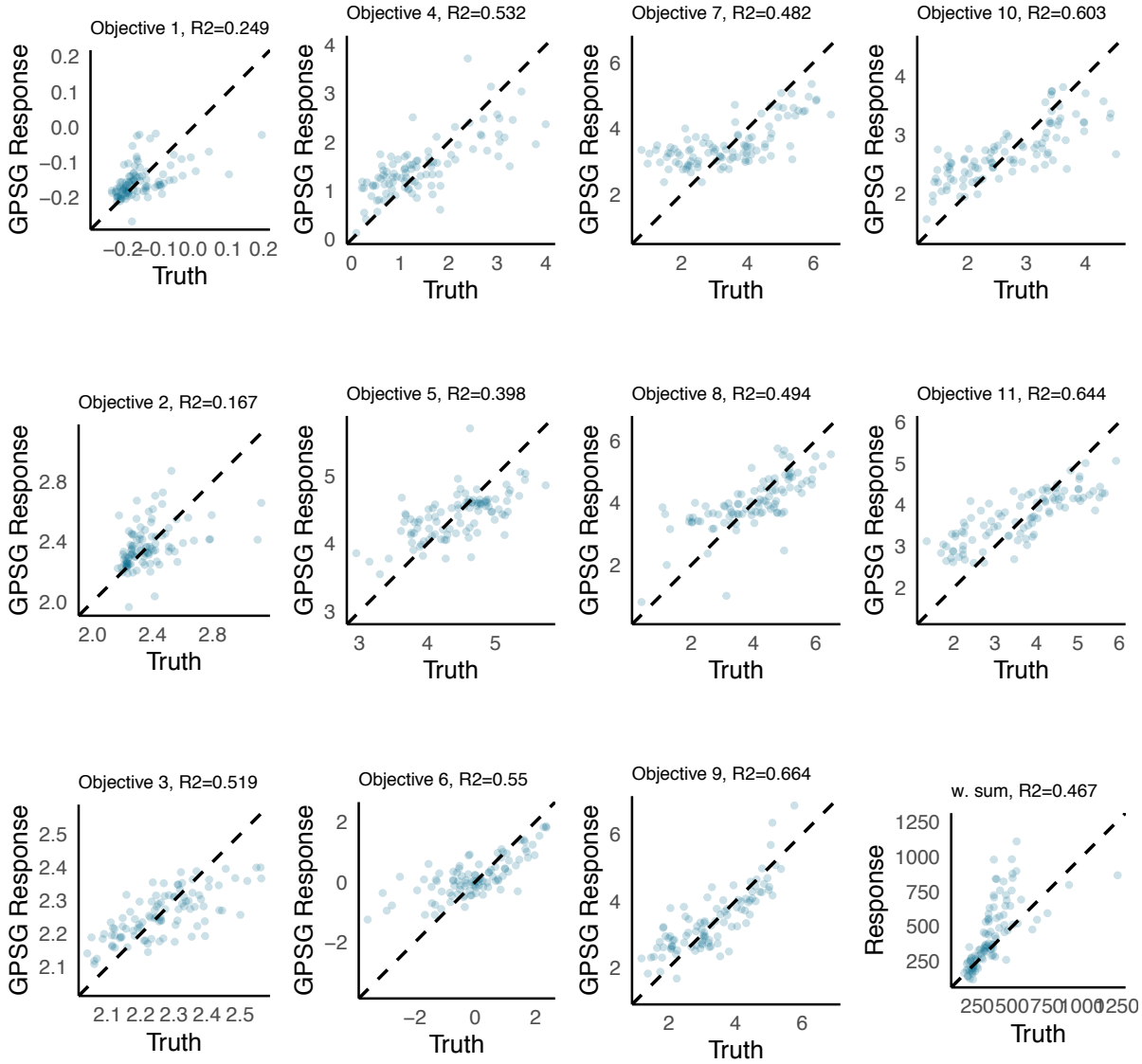


Figure S5. GPSG emulator performance. Emulator predictions vs true values on a holdout set compromising 10% of initial samples in iteration 23 (final iteration). w.sum is the weighted sum F, of the 11 objectives. Here, predictions are generated as the weighted sum of individual objective predictions.

5 ADAPTIVE SAMPLING: SELECTED POINTS

5.1 GP-BO

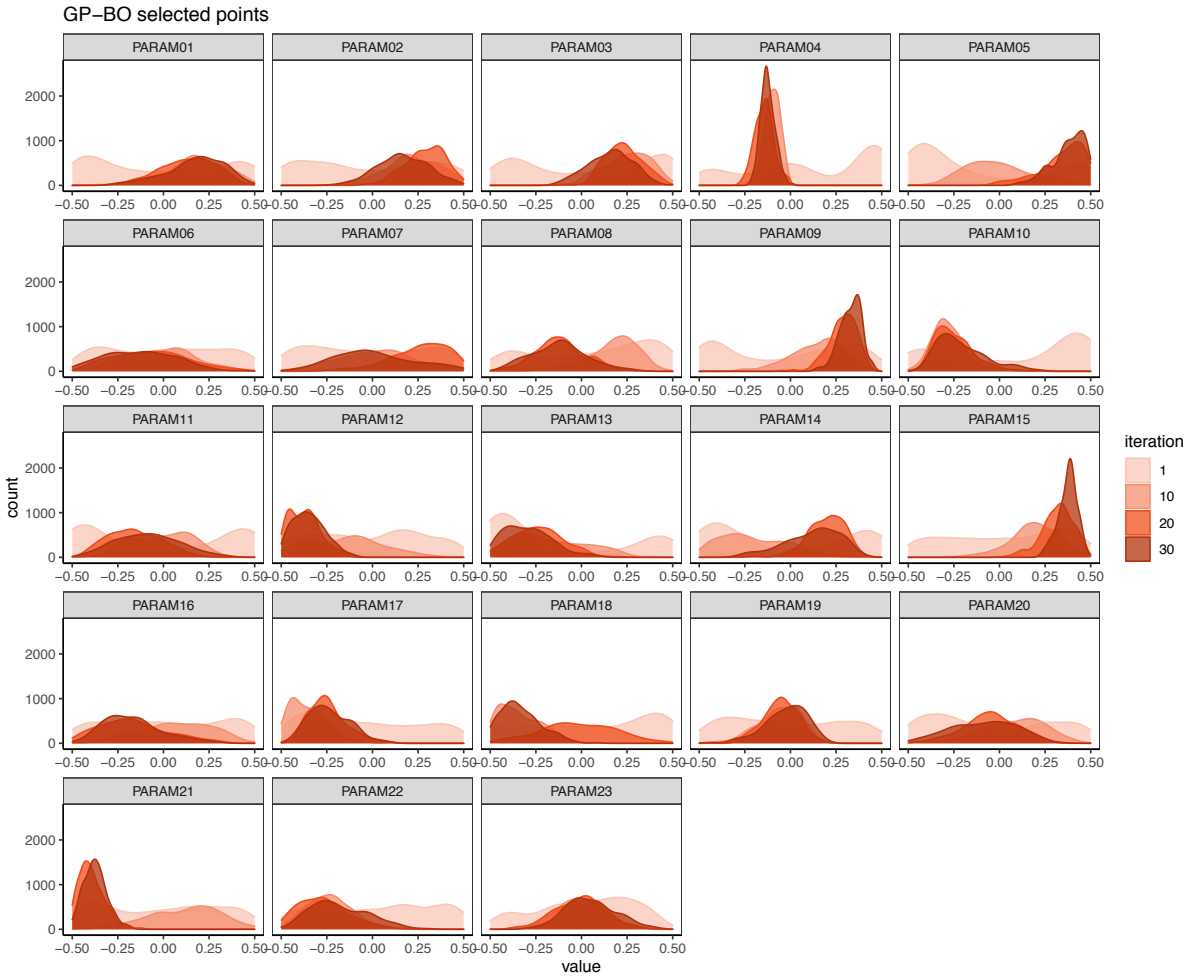


Figure S6. GP-BO sampling behavior. Values in each dimension of the points sampled during adaptive sampling of GP-BO algorithm in iterations 1,10, 20, and 30.

5.2 GPSG-BO

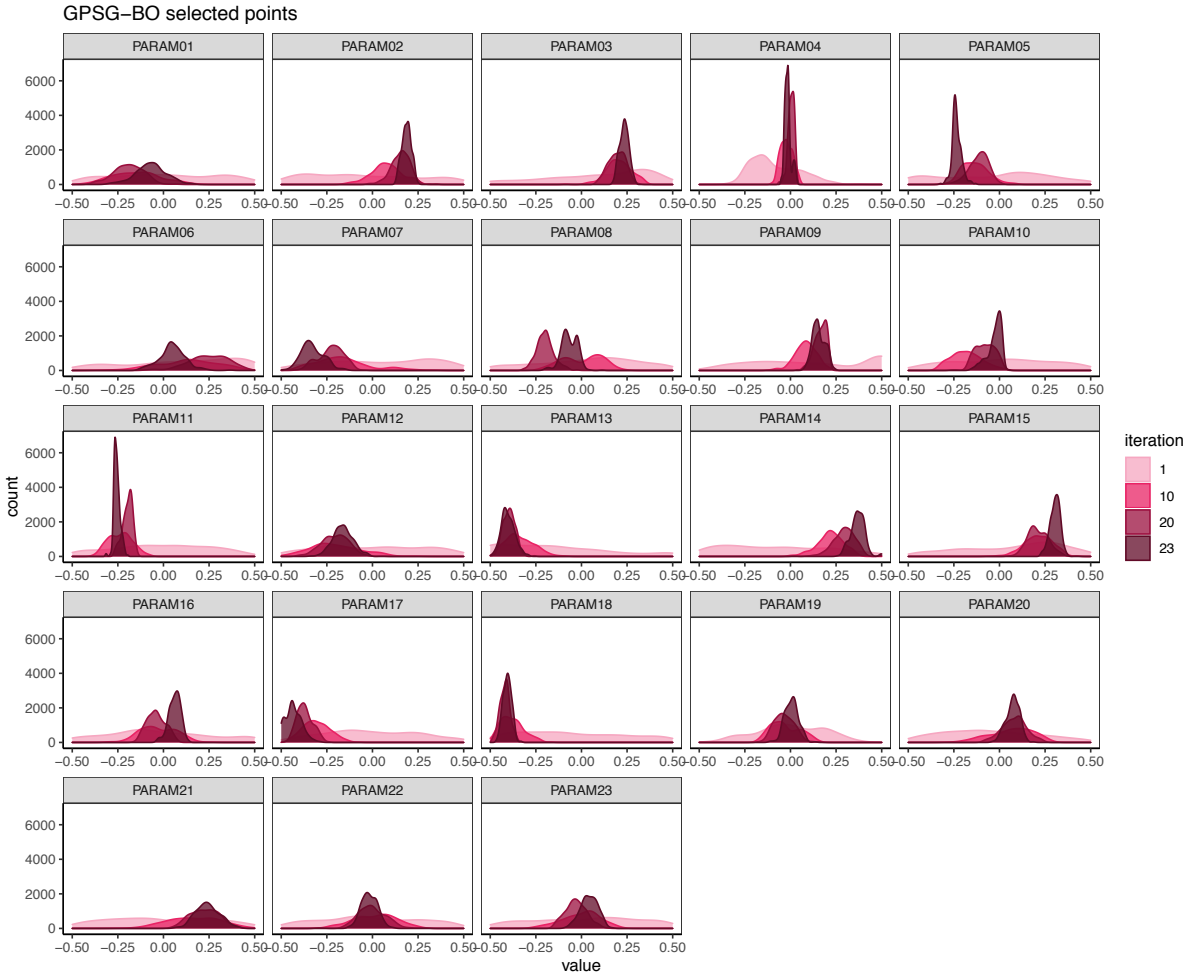


Figure S7. GPSG-BO sampling behavior. Values in each dimension of the points sampled during adaptive sampling of GPSG-BO algorithm in iterations 1,10, 20, and 23.

6 OPENMALARIA: FINAL SIMULATOR FIT

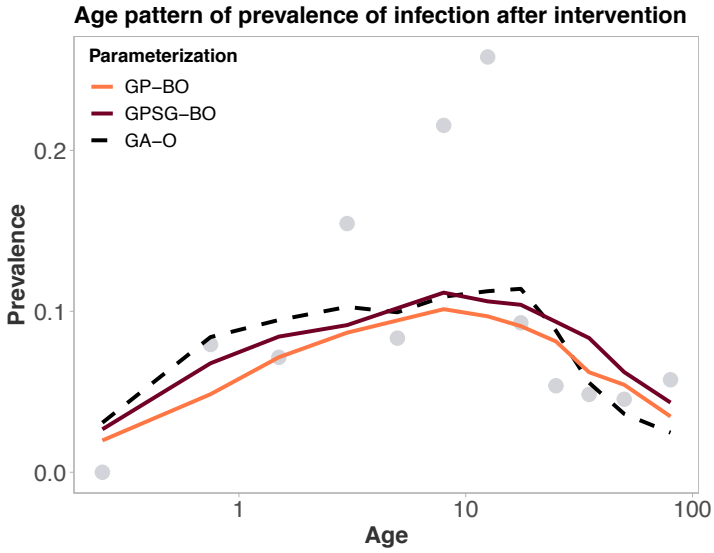


Figure S8. Objective 1: **Age pattern of prevalence in Matsari, Nigeria during the intervention.** Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

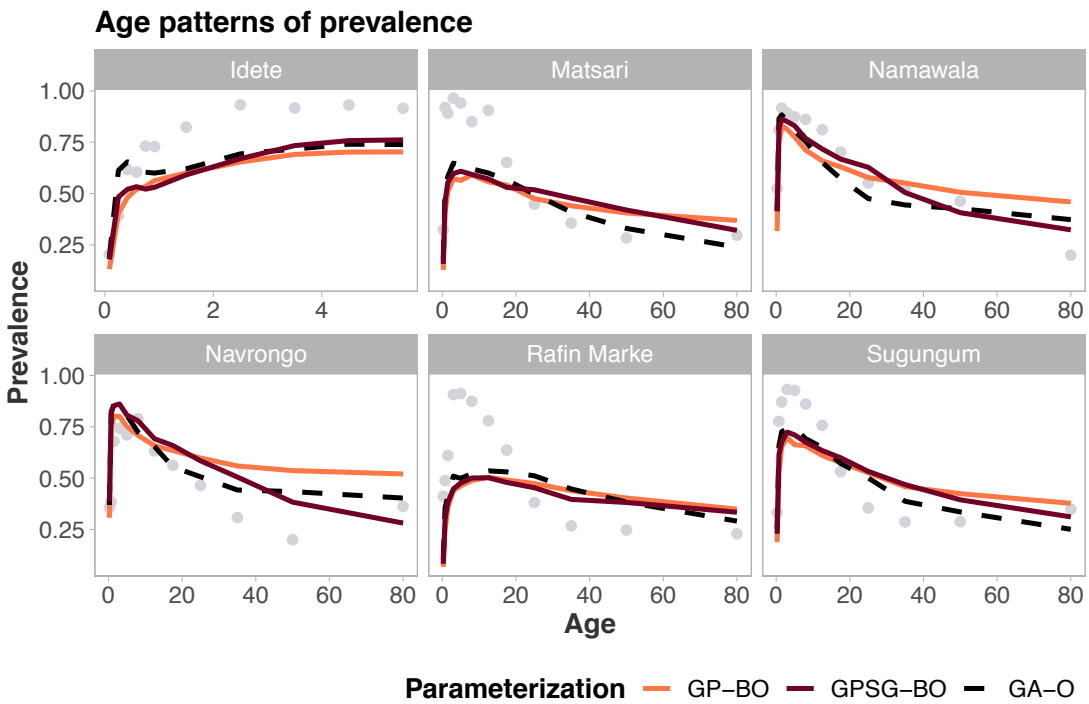


Figure S9. Objective 2: **Age pattern of prevalence.** Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

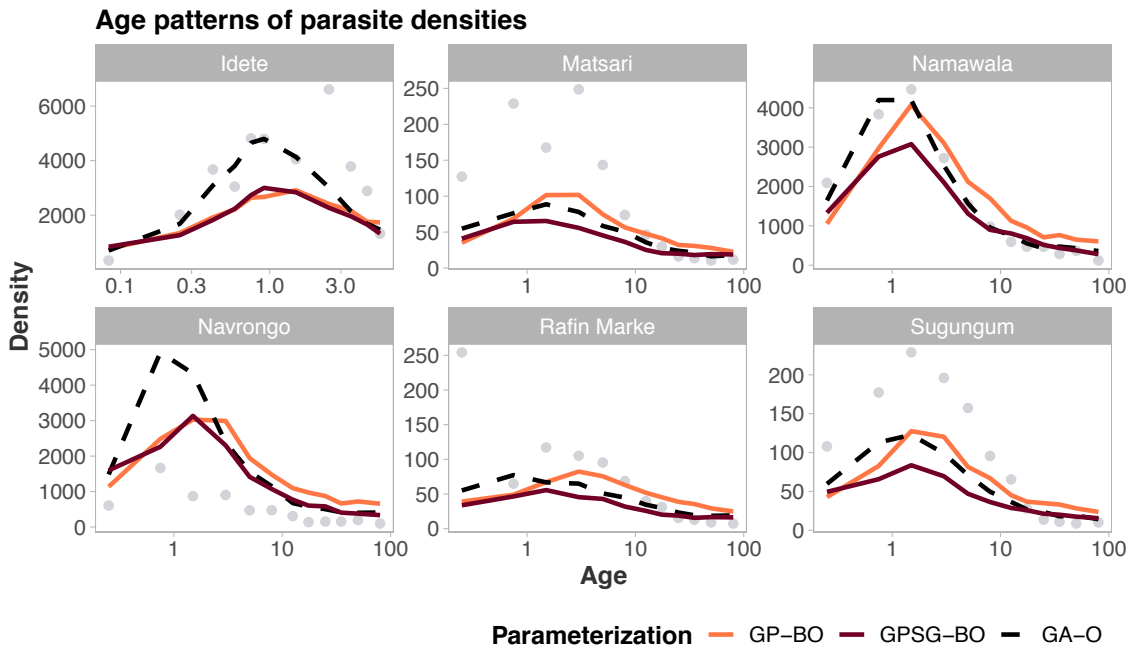


Figure S10. Objective 3: **Age pattern of parasite densities (geometric mean).** Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

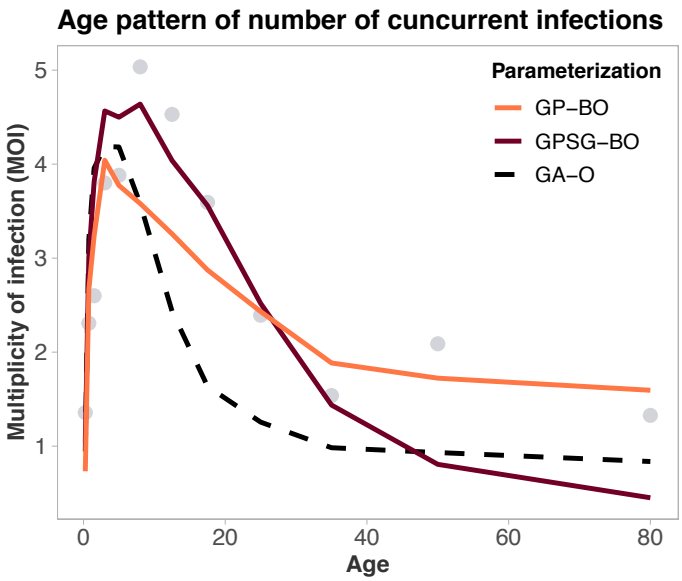


Figure S11. Objective 4: Age pattern of number of concurrent infections. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

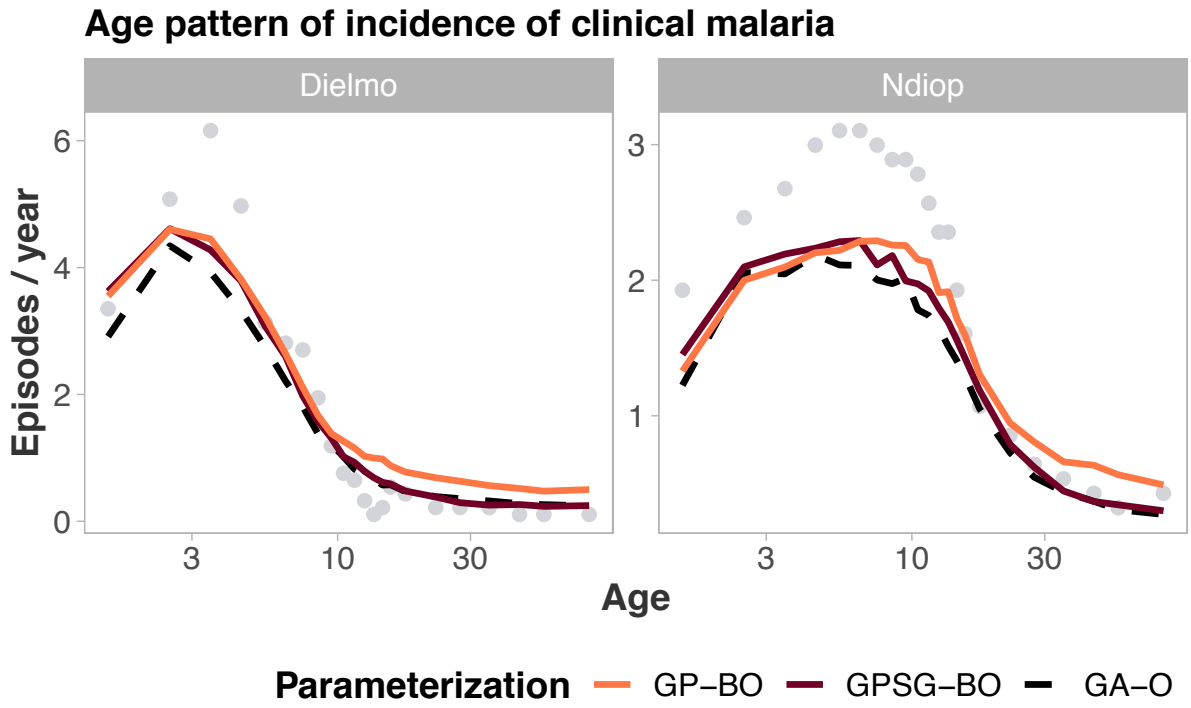


Figure S12. Objective 5: Age pattern of incidence of clinical malaria in Dielmo and Ndiop, Senegal. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

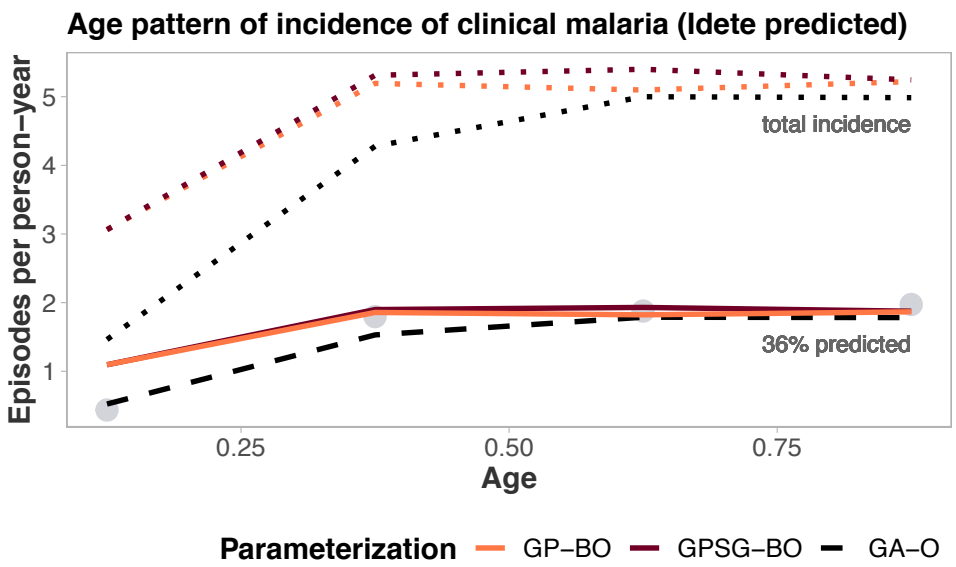


Figure S13. Objective 6: Age pattern of incidence of clinical malaria in Idete, Tanzania. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

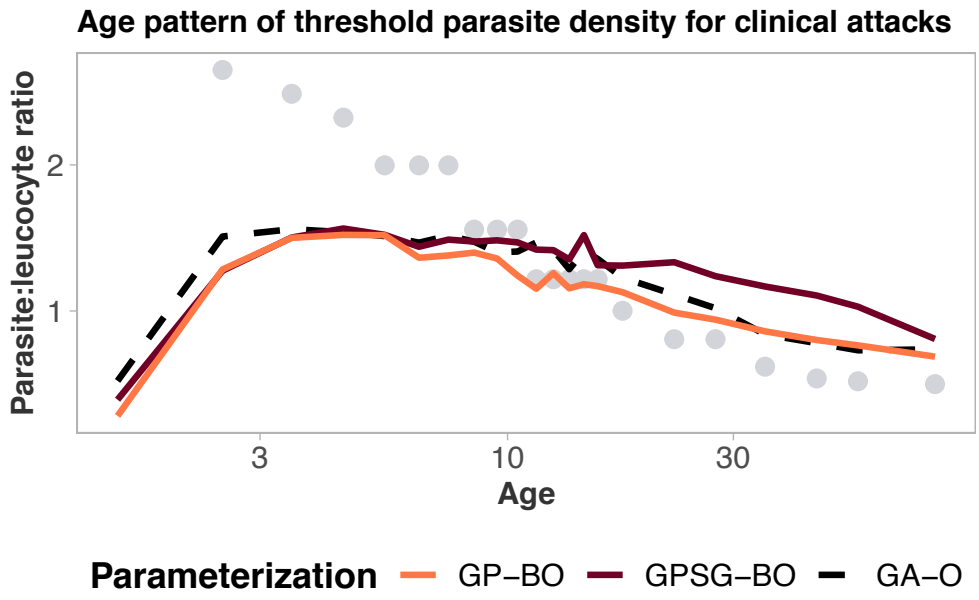


Figure S14. Objective 6. Age pattern of threshold parasite density for clinical attacks. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

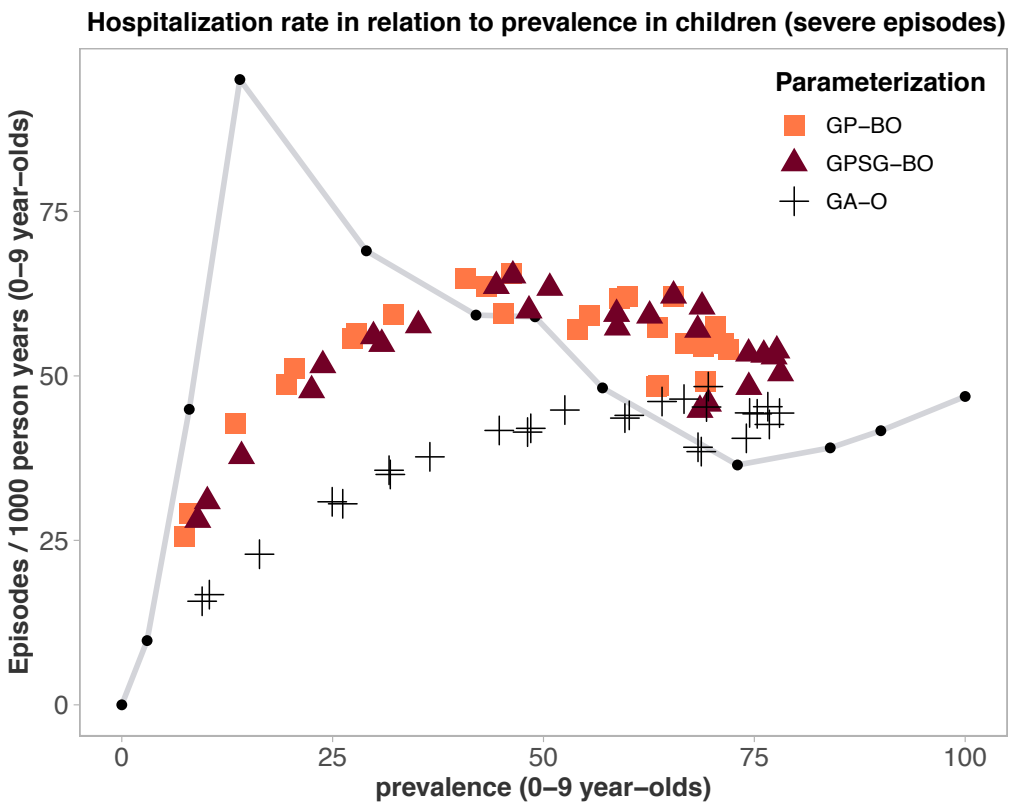


Figure S15. Objective 7: Hospitalization rate in relation to prevalence in children. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

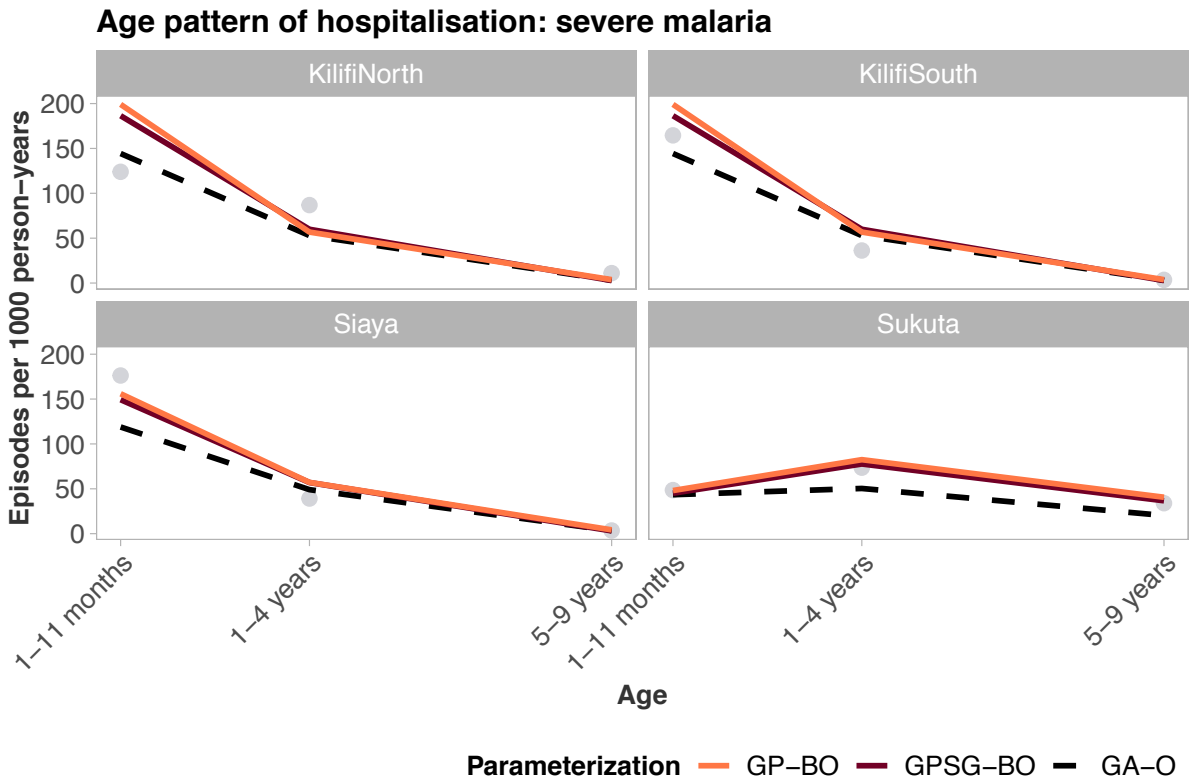


Figure S16. Objective 8. Age pattern of hospitalization. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

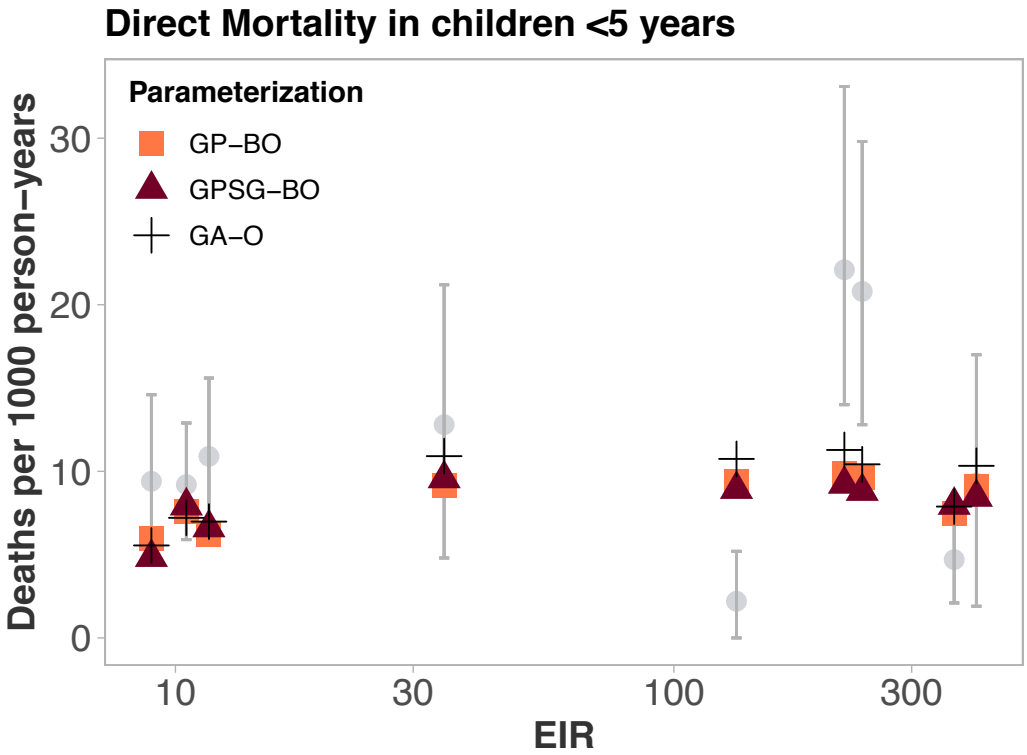


Figure S17. Objective 9: Direct mortality in children <5 years old. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

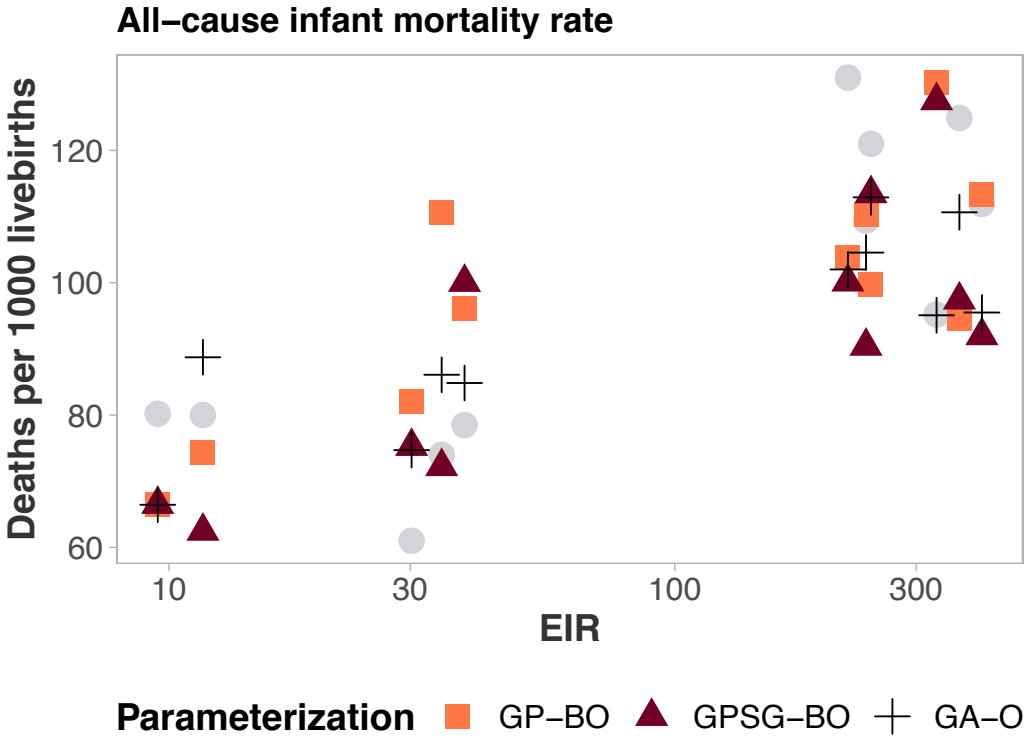
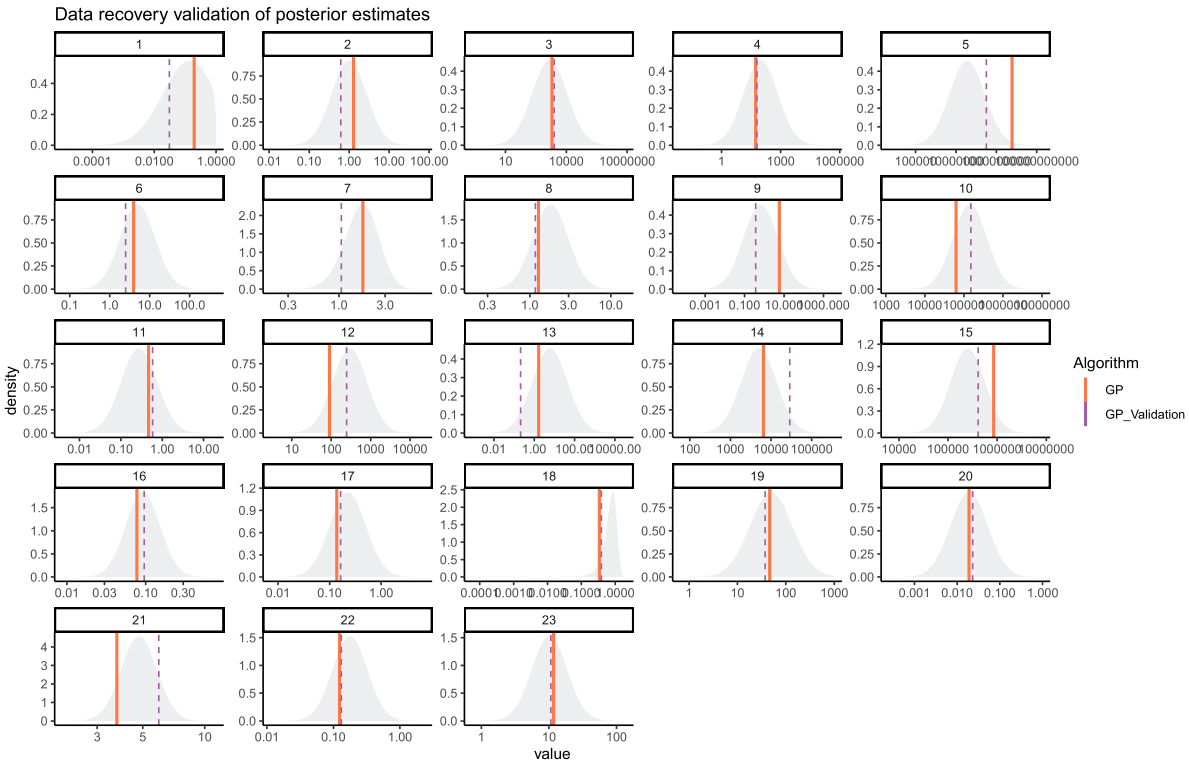


Figure S18. Objective 10: All-cause infant mortality rate. Final simulator fit using the parameter sets yielded using GP-BO and GPSG-BO compared to the previous parameterization (derived using optimization with a genetic algorithm, GA-O).

7 VALIDATION

A



B

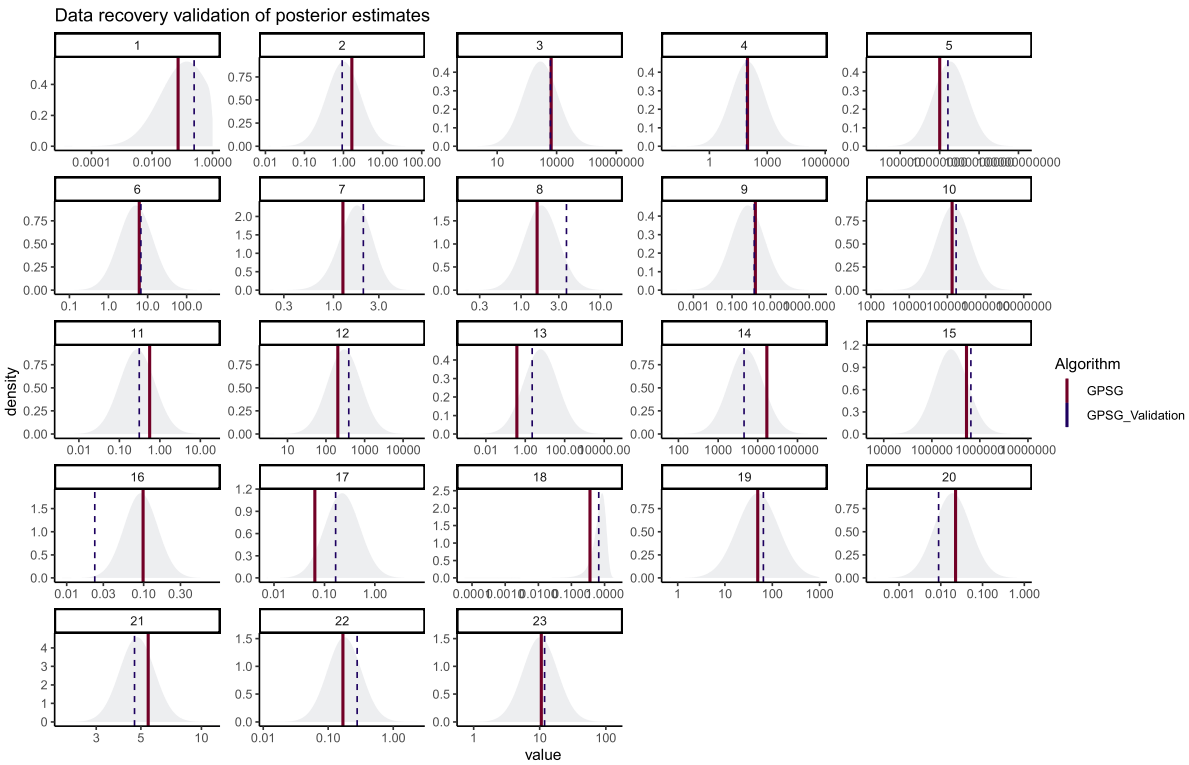


Figure S19. Data recovery validation of posterior estimates. Prior distributions of each parameter and parameter value identified by the optimization algorithm. The final parameter set was used to generate synthetic field data by simulating each of the 61 scenarios with the respective core

parameter sets. The simulation outputs were reformatted to match the original field data, generating a synthetic field data set. The optimization with both algorithms was repeated using this synthetic field data. The plot shows the best parameter values in each dimension identified at the end of the validation optimization compared to the values identified in the original optimization. The grey area shows the prior distribution. **A. GP-BO validation. B. GPSG-BO validation**

8 OPENMALARIA SIMULATED EPIDEMIOLOGY

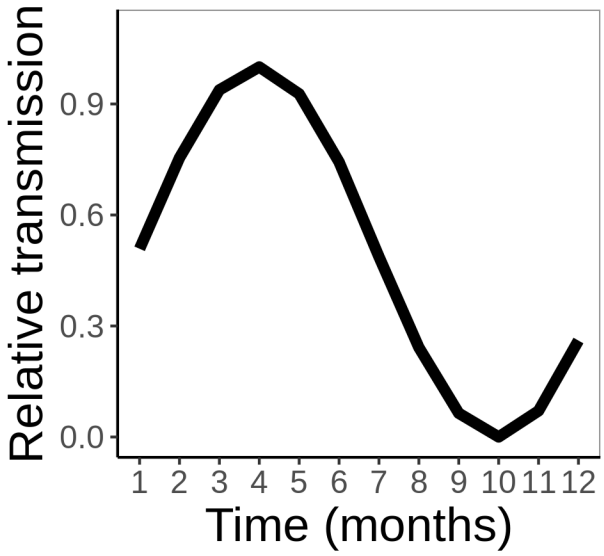


Figure S20. Seasonal pattern assumed for subsequent analyses. The monthly transmission intensity is equivalent to the annual transmission intensity (EIR) scaled by these values and forced to sum to the annual EIR.

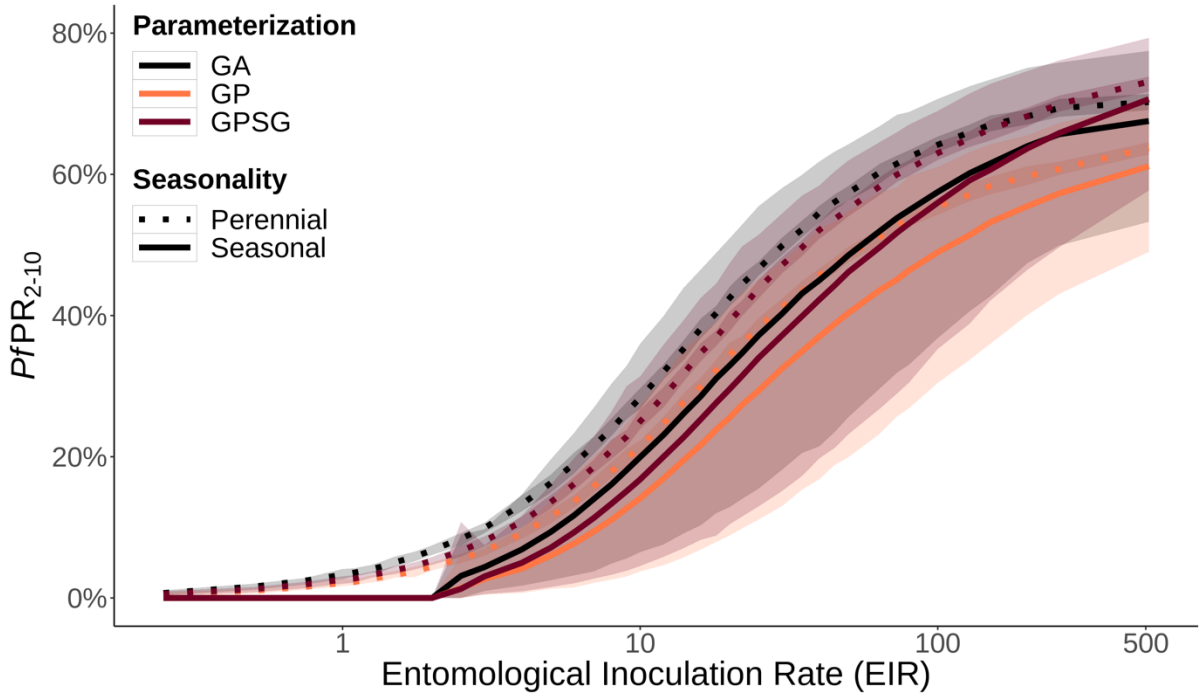


Figure S21. Relationship between EIR and PfPR2-10 under three parameterizations. Solid lines show medians and shaded regions show 95% credible intervals. EIR denotes the entomological inoculation rate.

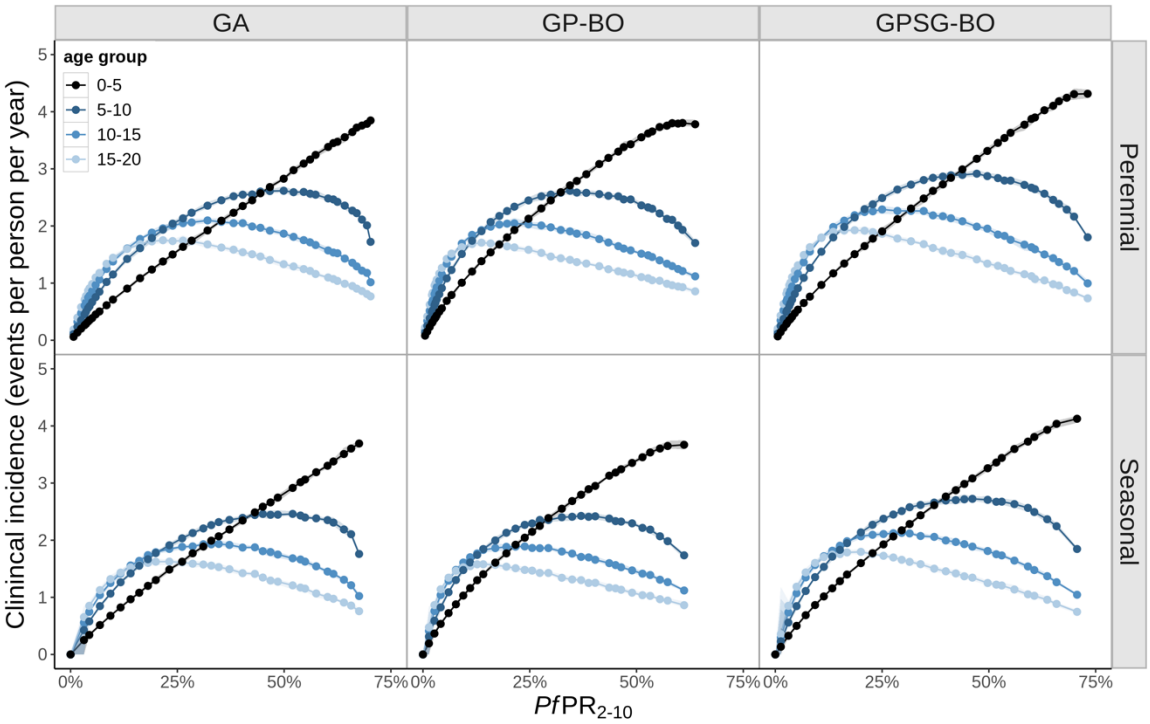


Figure S22. Yearly incidence of clinical (uncomplicated) malaria as a function of $PfPR_{2\text{-}10}$ displayed by parameterization and age group. Clinical incidence is presented in terms of the yearly number of events per person. We assume a probability of effective treatment within 14 days of uncomplicated malaria of 36%

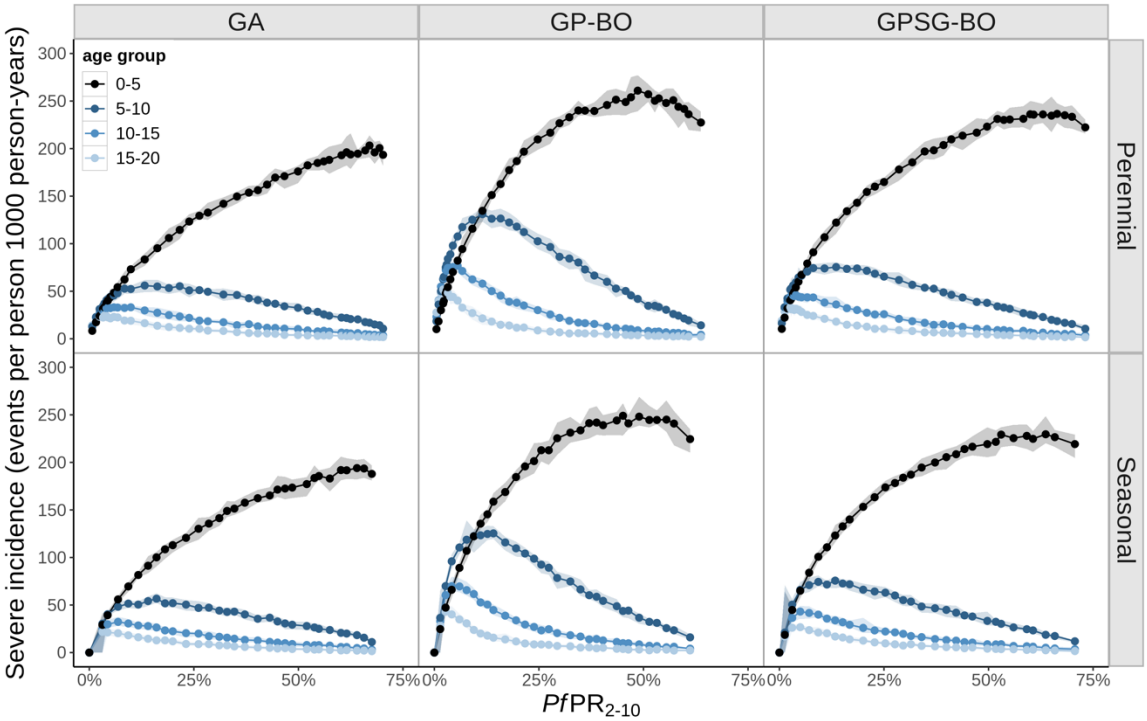


Figure S23. Yearly incidence of total severe malaria as a function of $PfPR_{2\text{-}10}$, displayed by parameterization and age group. Incidence is presented in terms of the yearly number of events in a population of 1000 individuals. It is assumed that 48% of severe malaria cases seek official care at a

heath care facility (hospital). We assume a probability of effective treatment within 14 days of uncomplicated malaria of 36%

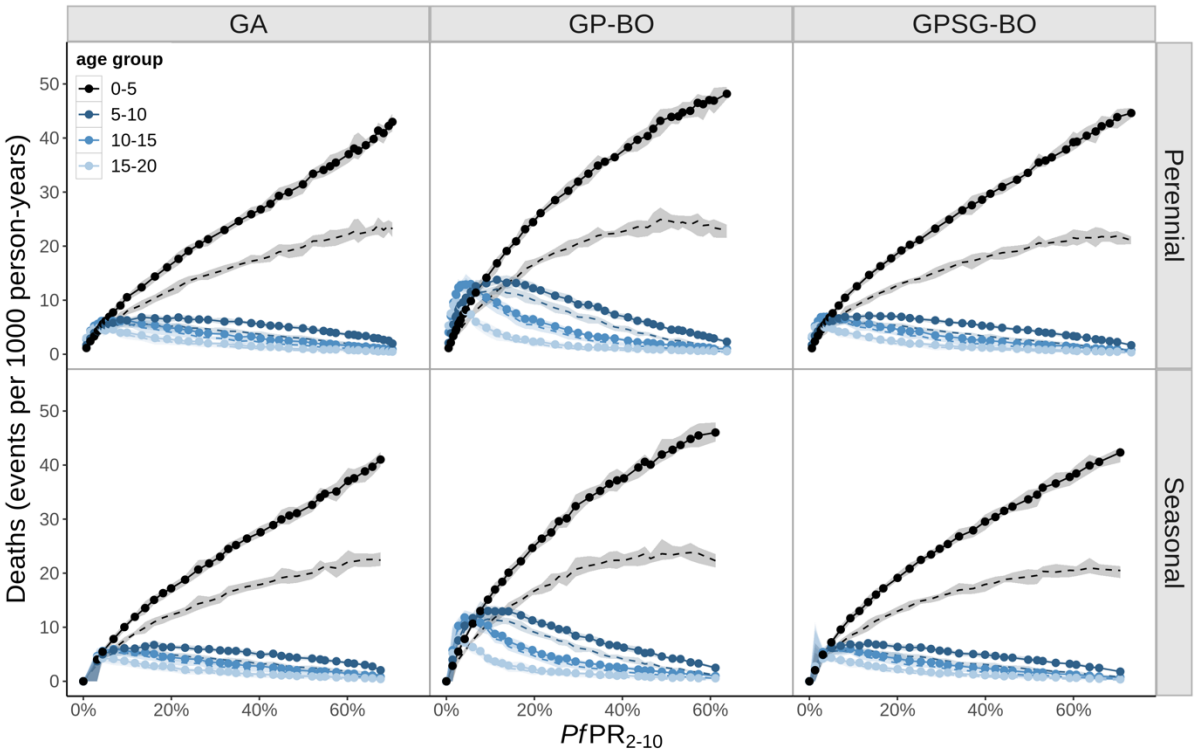


Figure S24. Yearly number of malaria-related deaths as a function of $PfPR_{2\text{-}10}$, displayed by parameterization and age group. Malaria mortality incidence is presented in terms of the yearly number of deaths in a population of 1000 individuals. For the OpenMalaria model both deaths directly attributed to malaria (dotted curve) and all deaths associated with malaria (including both deaths directly attributable to malaria and those associated with comorbidities) are shown (full line). See Box S1.2 for definitions of deaths attributable to malaria in the models

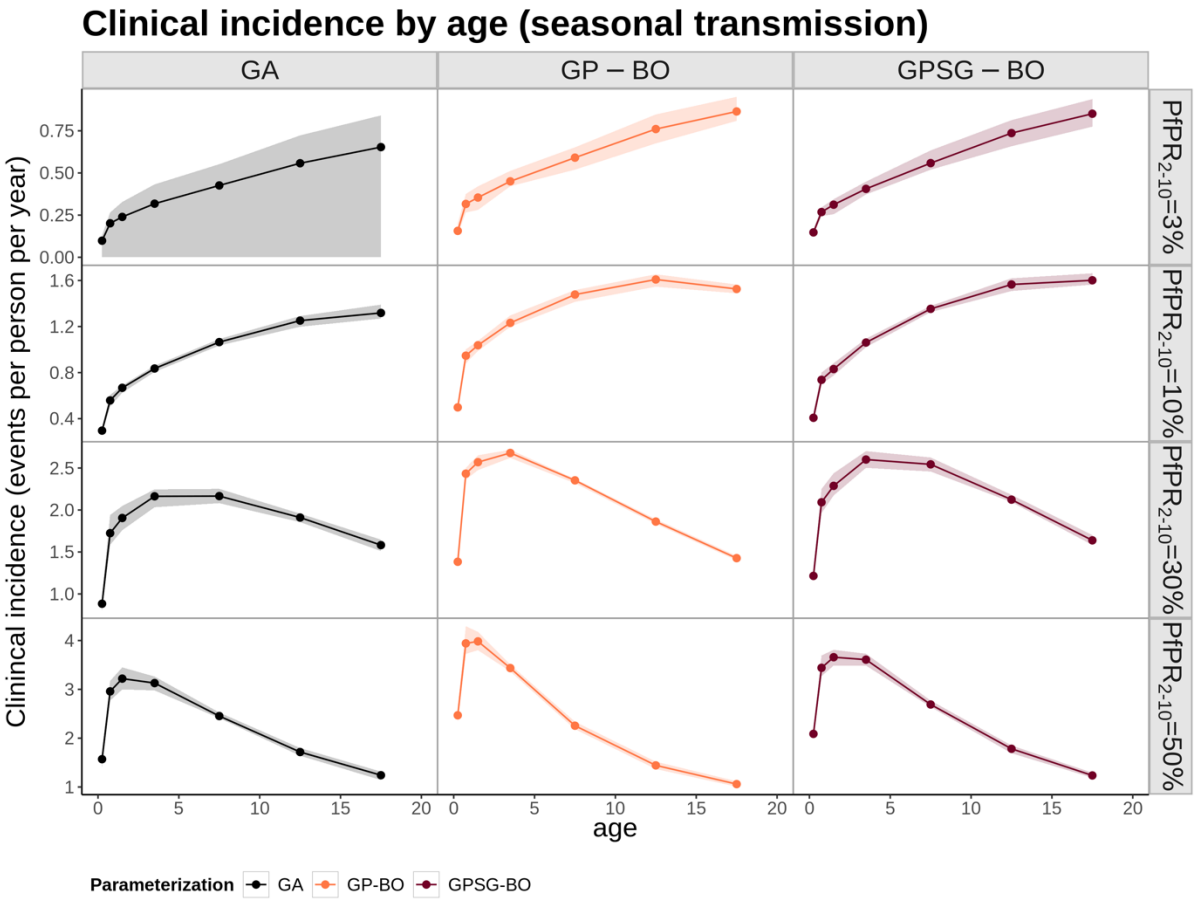


Figure S25. Yearly incidence of clinical malaria in a seasonal transmission setting as a function of age, displayed by transmission intensity ($PfPR_{2\text{-}10}$) and parameterization. Clinical incidence is presented in terms of the yearly number of events per person. The $PfPR_{2\text{-}10}$ categories include simulated prevalences of 2.5-3.5%, 9-10%, 28-32%, and 47-53% labeled as 3%, 10%, 30%, and 50%, respectively.

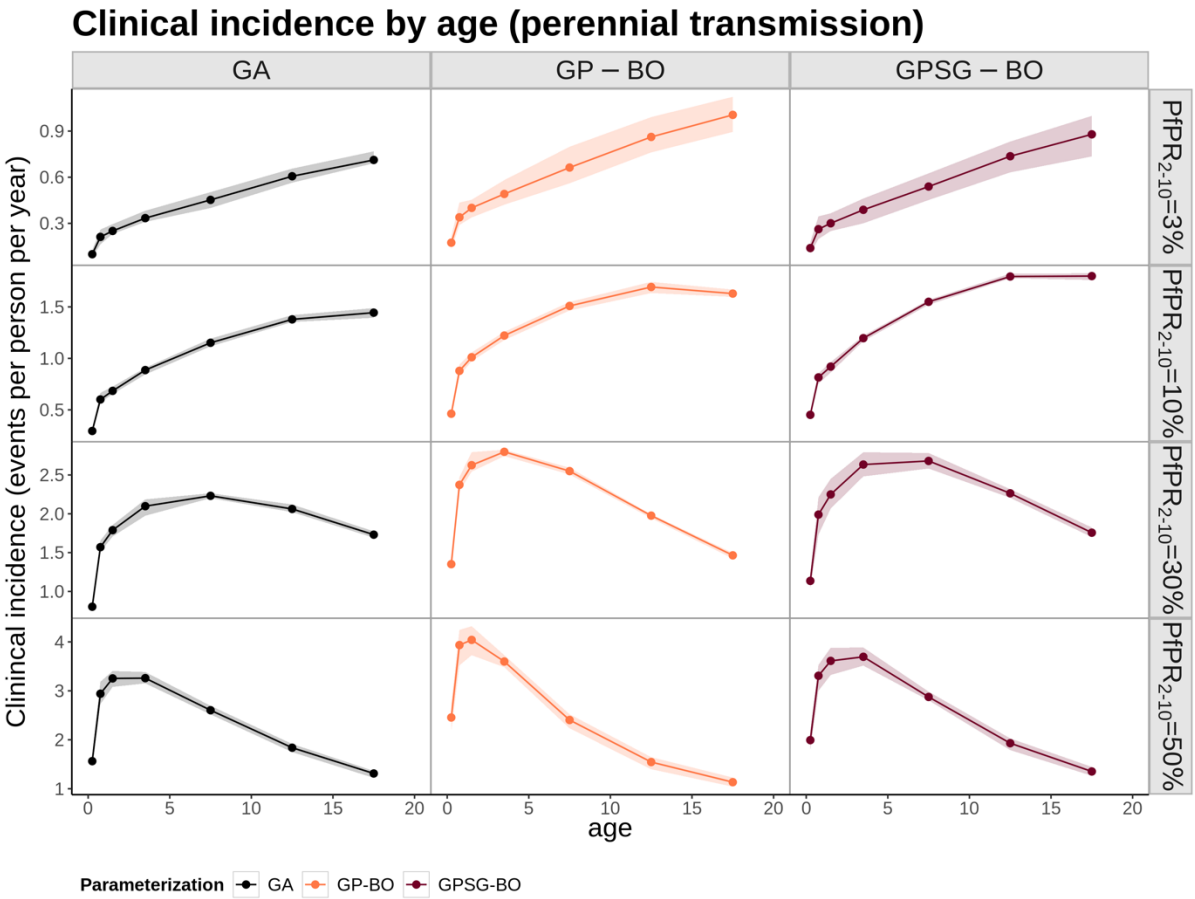


Figure S26. Yearly incidence of clinical malaria in a perennial transmission setting as a function of age, displayed by transmission intensity ($PfPR_{2\text{-}10}$) and parameterization. Clinical incidence is presented in terms of the yearly number of events per person. The $PfPR_{2\text{-}10}$ categories include simulated prevalences of 2.5-3.5%, 9-10%, 28-32%, and 47-53% labeled as 3%, 10%, 30%, and 50%, respectively

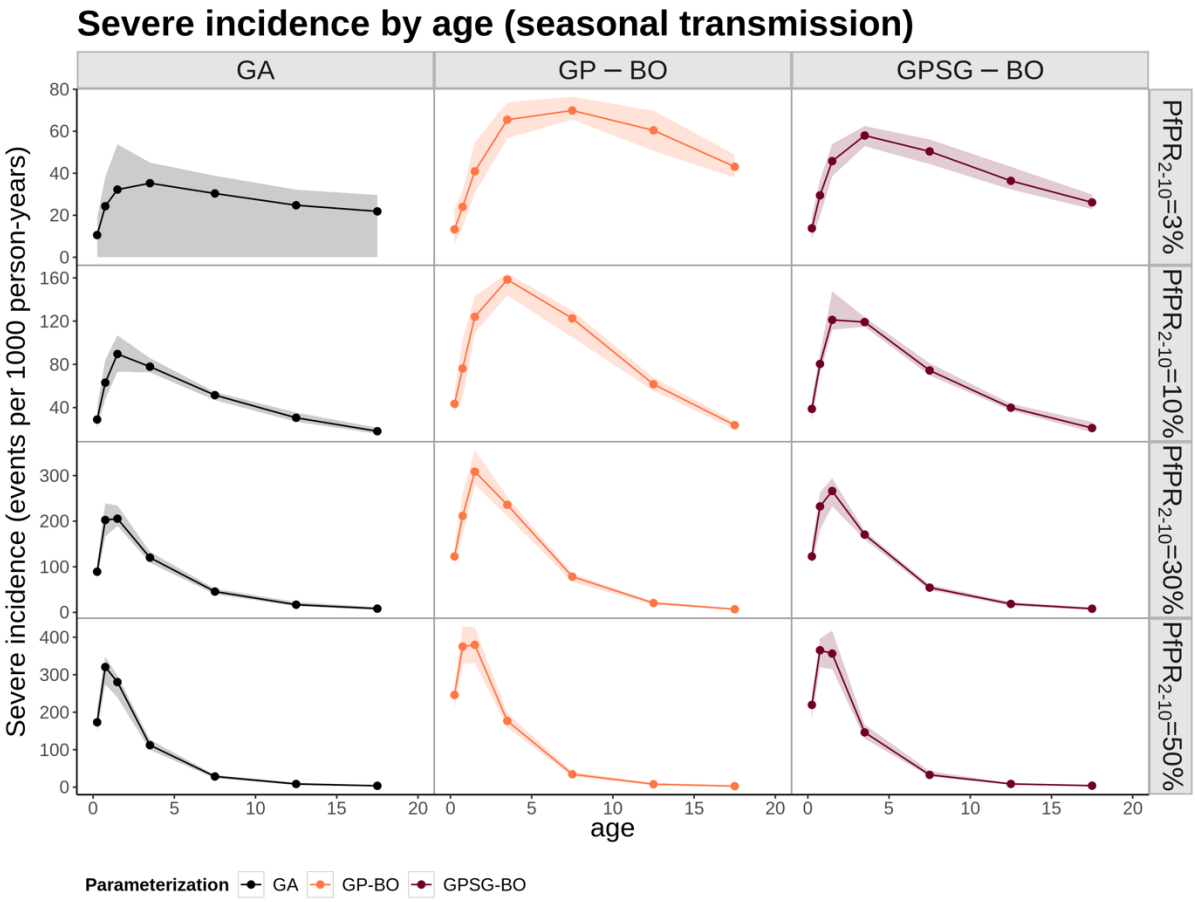


Figure S27. Yearly incidence of total severe malaria in a seasonal transmission setting as a function of age, displayed by transmission intensity ($PfPR_{2\text{-}10}$) and parameterization. Incidence is presented in terms of the yearly number of events per 1000 person-years. It is assumed that 48% of severe malaria cases seek official care at a heath care facility (hospital). The $PfPR_{2\text{-}10}$ categories include simulated prevalences of 2.5-3.5%, 9-10%, 28-32%, and 47-53% labeled as 3%, 10%, 30%, and 50%, respectively

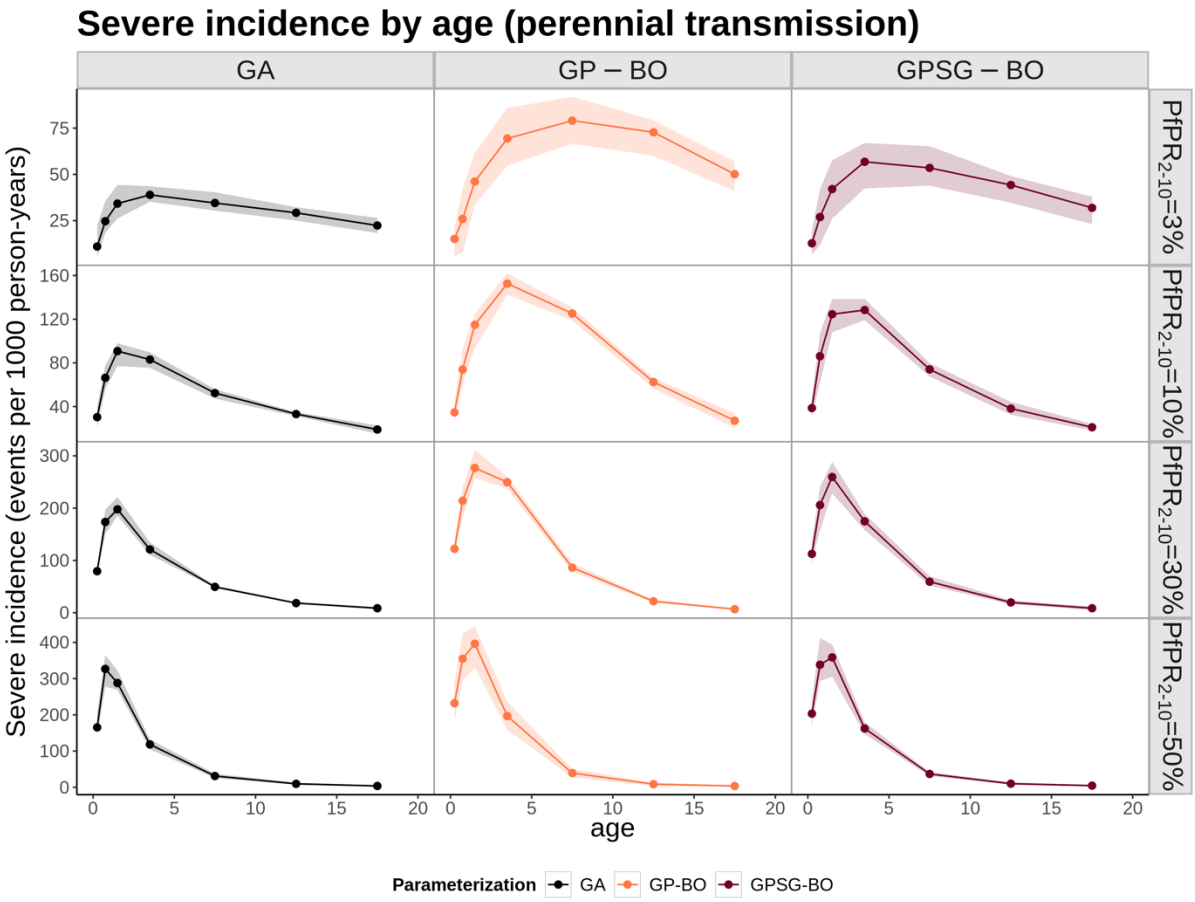


Figure S28. Yearly incidence of total severe malaria in a perennial transmission setting as a function of age, displayed by transmission intensity ($PfPR_{2\text{-}10}$) and parameterization. Incidence is presented in terms of the yearly number of events per 1000 person-years. It is assumed that 48% of severe malaria cases seek official care at a heath care facility (hospital). The $PfPR_{2\text{-}10}$ categories include simulated prevalences of 2.5-3.5%, 9-10%, 28-32%, and 47-53% labeled as 3%, 10%, 30%, and 50%, respectively

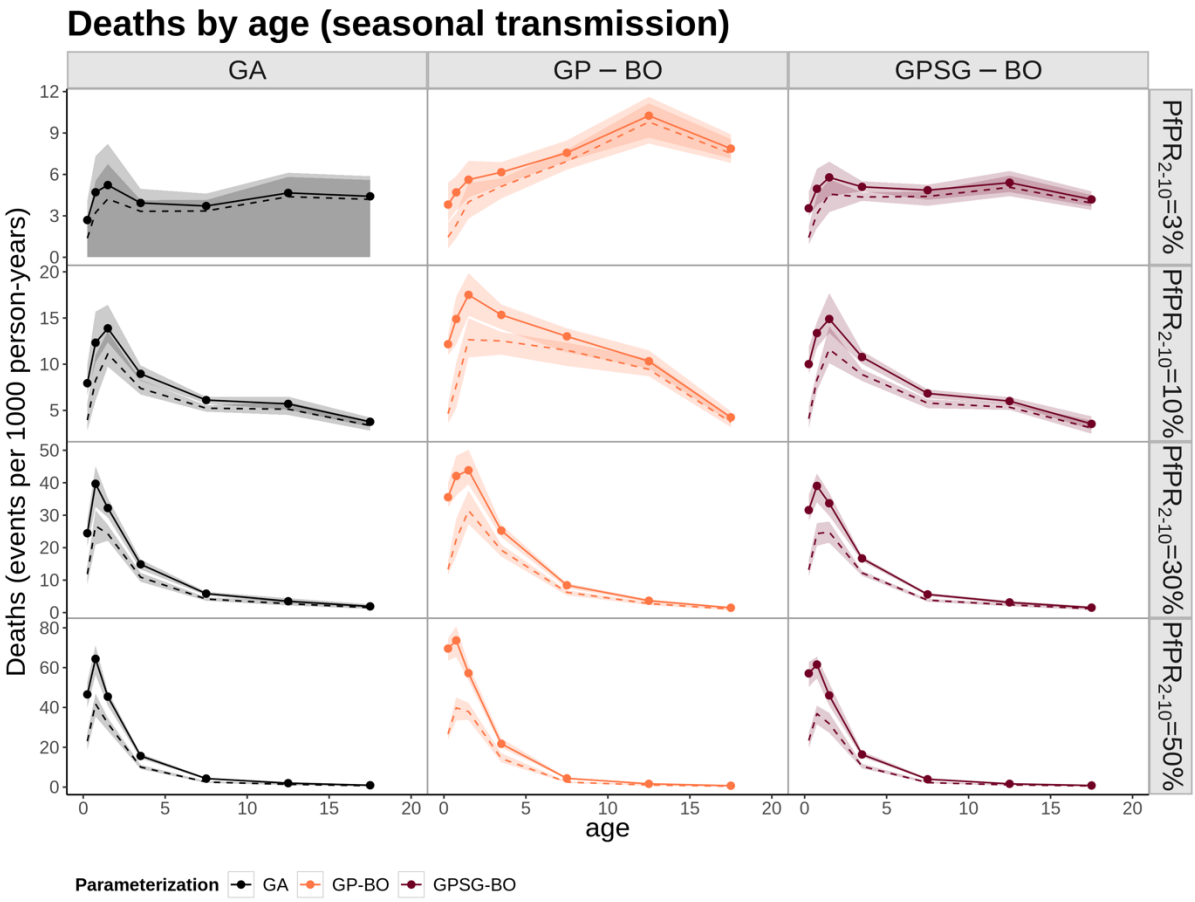


Figure S29. Yearly incidence of malaria-related deaths in a seasonal transmission setting as a function of age, displayed by transmission intensity ($PfPR_{2\text{-}10}$) and parameterization. Malaria mortality incidence is presented in terms of the yearly number of deaths in a population of 1000 individuals. The dashed estimates represent direct malaria deaths, and the solid all malaria deaths (including those attributable to co-morbidities). The $PfPR_{2\text{-}10}$ categories include simulated prevalences of 2.5-3.5%, 9-10%, 28-32%, and 47-53% labeled as 3%, 10%, 30%, and 50%, respectively

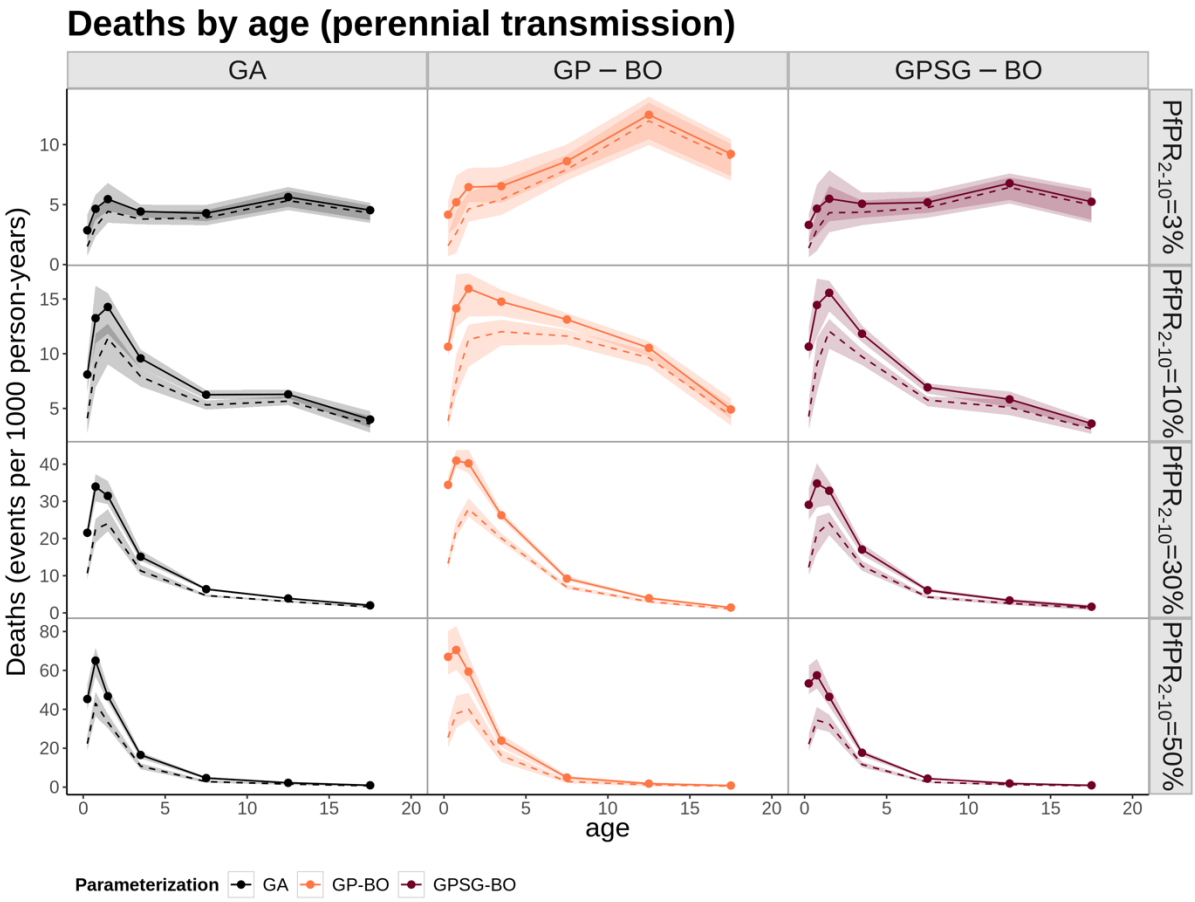


Figure S30. Yearly incidence of malaria-related deaths in a perennial transmission setting as a function of age, displayed by transmission intensity ($PfPR_{2\text{-}10}$) and parameterization. Malaria mortality incidence is presented in terms of the yearly number of deaths in a population of 1000 individuals. The dashed estimates represent direct malaria deaths, and the solid all malaria deaths (including those attributable to co-morbidities). The $PfPR_{2\text{-}10}$ categories include simulated prevalences of 2.5-3.5%, 9-10%, 28-32%, and 47-53% labeled as 3%, 10%, 30%, and 50%, respectively

9 LOG PRIOR DISTRIBUTIONS

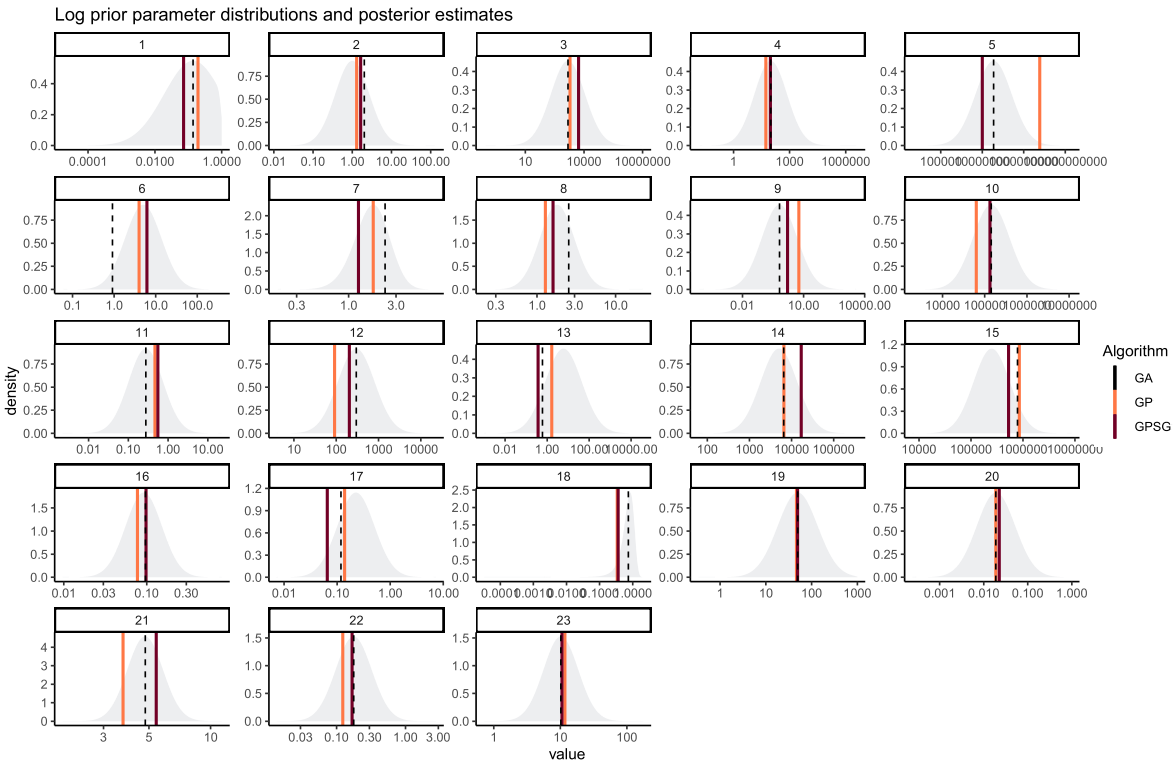


Figure S31. Log prior distributions and final posterior estimates. Prior distributions of each parameter and final parameter values identified by each optimization algorithm (GP-BO and GPSG-BO) and compared to the current parameterization (derived using a genetic algorithm, GA).

10 RANGER IMPORTANCE

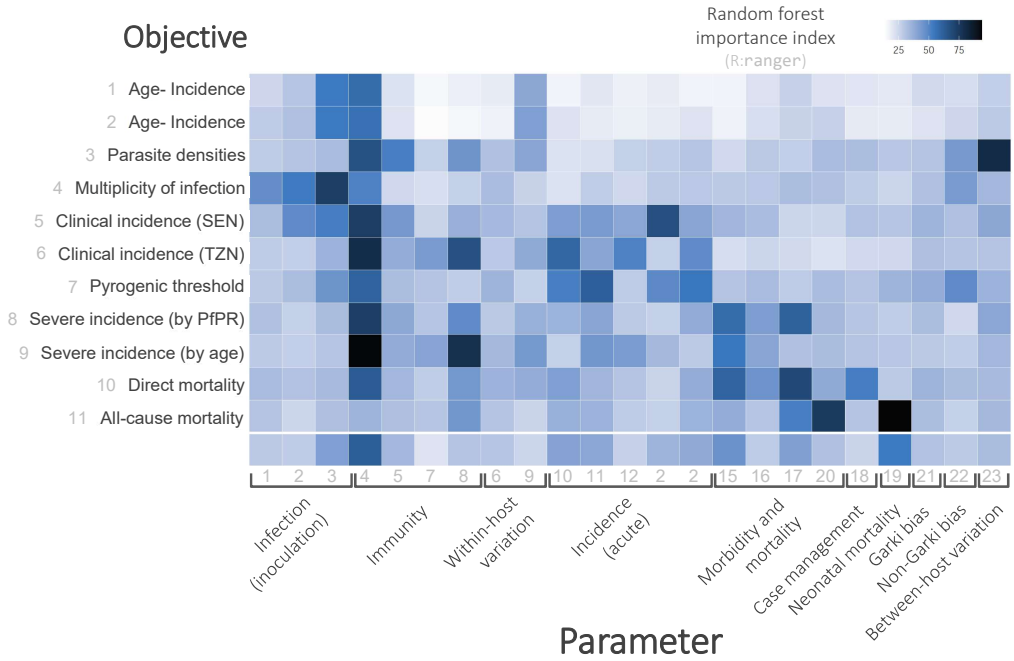


Figure S32. Log prior distributions and final posterior estimates. Prior distributions of each parameter and final parameter values identified by each optimization algorithm (GP-BO and GPSG-BO) and compared to the current parameterization (derived using a genetic algorithm, GA).