Improved prediction accuracy for disease risk mapping using Gaussian Process stacked generalisation

Samir Bhatt[1,*], Ewan Cameron[2], Seth R Flaxman[4], Daniel J Weiss[2], David L Smith[3], and Peter W Gething[2]

[1] *Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK*
[2] *Oxford Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK*
[3] *Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington 98121, USA*
[4] *Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK*
[*] *Corresponding author: bhattsamir@gmail.com*

Abstract

Maps of infectious disease—charting spatial variations in the force of infection, degree of endemicity, and the burden on human health—provide an essential evidence base to support planning towards global health targets. Contemporary disease mapping efforts have embraced statistical modelling approaches to properly acknowledge uncertainties in both the available measurements and their spatial interpolation. The most common such approach is Gaussian process regression, a mathematical framework comprised of two components: a mean function harnessing the predictive power of multiple independent variables, and a covariance function yielding spatio-temporal shrinkage against residual variation from the mean. Though many techniques have been developed to improve the flexibility and fitting of the covariance function, models for the mean function have typically been restricted to simple linear terms. For infectious diseases, known to be driven by complex interactions between environmental and socio-economic factors, improved modelling of the mean function can greatly boost predictive power. Here we present an ensemble approach based on stacked generalisation that allows for multiple non-linear algorithmic mean functions to be jointly embedded within the Gaussian process framework. We apply this method to mapping *Plasmodium falciparum* prevalence data in Sub-Saharan Africa and show that the generalised ensemble approach markedly out-performs any individual method.

Author Summary

Infectious disease mapping provides a powerful synthesis of evidence in an effective, visually condensed form. With the advent of new web-based data sources and systematic data collection in the form of cross-sectional surveys and health facility reporting, there is high demand for accurate methods to predict spatial maps. The primary technique used in spatial mapping is known as Gaussian process regression (GPR). GPR is a flexible stochastic model that allows the modelling of disease-driving factors such as the environment while also capturing unknown residual spatial correlations in the data. We introduce a method that blends state-of-the-art machine learning methods with GPR to produce a model that substantially out-performs other methods commonly used in disease mapping. The utility of this new approach also extends far beyond just mapping and can be used for general machine learning applications across computational biology, including Bayesian optimisation and mechanistic modelling.

Introduction

Infectious disease mapping with model-based geostatistics [1] can provide a powerful synthesis of the available evidence base to assist surveillance systems and support progress towards global health targets, revealing the geographical bounds of disease occurrence and the spatial patterns of transmission intensity and clinical burden. A recent review found that out of 174 infectious diseases with a strong rational for mapping, only 7 (4%) have thus far been comprehensively mapped [2]. The primary factor impeding progress is a lack of accurate, population representative, geopositioned data. In recent years this has begun to change as increasing volumes of spatially referenced data are collected from both cross-sectional household surveys and web-based data sources (e.g. Health Map [3]), bringing new opportunities for scaling up the global mapping of diseases. Alongside this surge in new data, novel statistical methods are needed that can generalise to new data accurately while remaining computationally tractable on large datasets. In this paper we will introduce one such method designed with these aims in mind.

Owing to both a long history of published research in the field and a widespread appreciation amongst endemic countries for the value of cross-sectional household surveys as guides to

intervention planning, malaria is an example of a disease that *has* been comprehensively mapped. Over the past decade, volumes of publicly-available malaria prevalence data—defined as the proportion of parasite positive individuals in a sample—have reached sufficiency to allow for detailed spatio-temporal mapping [4]. From a statistical perspective, the methodological mainstay of these malaria prevalence mapping efforts has been Gaussian process regression [5–8]. Gaussian processes are a flexible semi-parametric regression technique defined entirely through a mean function, $\mu(\cdot)$, and a covariance function, $k(\cdot, \cdot)$. The mean function models an underlying trend, such as the effect of environmental/socio-economic factors, while the covariance function applies Bayesian shrinkage to residual variation from the mean such that points close to each other in space and time tend towards similar values. The resulting ability of Gaussian processes to strike a parsimonious balance in the weighting of explained and unexplained spatio-temporal variation has led to their near exclusive use in contemporary studies of the geography of malaria prevalence [1, 4, 7–10].

Outside of disease mapping, Gaussian processes have been used for numerous applications in machine learning, including regression [1, 5, 6], classification [5], and optimisation [11]; their popularity leading to the development of efficient computational techniques and statistical parametrisations. A key challenge for the implementation of Gaussian process models arises in the statistical learning (or inference) of the underlying parameters controlling the chosen mean and covariance functions. Learning is typically performed using Markov Chain Monte Carlo (MCMC) or by maximizing the marginal likelihood [5], both of which are made computationally demanding by the need to compute large matrix inverses returned by the covariance function. The complexity of this inverse operation is $\mathcal{O}(n^3)$ in computation and $\mathcal{O}(n^2)$ in storage in the naive case [5], which imposes practical limits on data sizes [5]. MCMC techniques may be further confounded by mixing problems in the Markov chains. These challenges have necessitated the use of highly efficient MCMC methods, such as Hamiltonian MCMC [12] or posterior approximation approaches, such as the integrated nested Laplace approximation [13], expectation propagation [5,14,15], and variational inference [16,17]. Additionally many frequentist approaches have been developed including matrix free [18] and primal learning approaches [19]. Many of these methods adopt finite dimensional representations of the covariance function yielding sparse precision matrices, either by specifying a fully independent training conditional (FITC)

structure [20] or by identifying a Gaussian Markov Random Field (GMRF) approximation to the continuous process [21].

Alongside these improved methods for inference, recent research has focussed on model development to increase the flexibility and diversity of parametrisations for the covariance function, with new techniques utilising solutions to stochastic partial differential equations (allowing for easy extensions to non-stationary and anisotropic forms [21]), the combination of kernels additively and multiplicatively [22], and various spectral representations [23].

One aspect of Gaussian processes that has remained largely neglected is the mean function which is often—and indeed with justification in some settings—simply set to zero and ignored. However, in the context of disease mapping, where the biological phenomena are driven by a complex interplay of environmental and socioeconomic factors [24], the mean plays a central role in improving the predictive performance of Gaussian process models. Furthermore, it has also been shown that using a well-defined mean function can allow for simpler covariance functions (and hence simpler, scalable inference techniques) [25].

The steady growth of remotely-sensed data with incredible spatio-temporal richness [24] combined with well-developed biological models [26] has meant that there is a rich suite of environmental and socio-economic covariates currently available. In previous malaria mapping efforts these covariates have been modelled as simple linear predictors [7–9] that fail to capture complex non-linearities and interactions, leading to a reduced overall predictive performance. Extensive covariate engineering can be performed by introducing large sets of non-linear and interacting transforms of the covariates, but this brute force combinatorial problem quickly becomes computationally inefficient [4, 24].

In the field of machine learning and data science there has been great success with algorithmic approaches that neglect the covariance and focus on learning from the covariates alone [27, 28]. These include tree based algorithms such as boosting [29] and random forests [30], generalized additive spline models [31, 32], multivariate adaptive regression splines [33], and regularized regression models [34]. The success of these methods is grounded in their ability to manipulate the bias-variance trade-off [35], capture interacting non-linear effects, and perform automatic covariate selection. The technical challenges of hierarchically embedding these algorithmic

methods within the Gaussian process framework are forbidding and many of the approximation methods that make Gaussian process models computationally tractable would struggle with their inclusion. Furthermore, it is unclear which of these approaches would best characterize the mean function when applied across different diseases and settings. In this paper we propose a simplified embedding method based on stacked generalisation [36, 37] that focuses on improving the mean function of a Gaussian process, thereby allowing for substantial improvements in the predictive accuracy beyond what has been achieved in the past.

Methods

Gaussian process regression

We define our response, $\mathbf{y}_{s,t} = \{y_{(s,t)[1]}, ..., y_{(s,t)[n]}\}$, as a vector of n empirical logit transformed malaria prevalence surveys at location-time pairs, $(s,t)[i]$, with $\mathbf{X}_{s,t} = \{(\mathbf{x}_{1:m})[1], \ldots, (\mathbf{x}_{1:m})[n]\}$ denoting a corresponding $n \times m$ design matrix of m covariates (see section [Data, Covariates and Experimental Design] below). The likelihood of the observed response is $\mathbb{P}(\mathbf{y}_{s,t}|\mathbf{f}_{s,t}, \mathbf{X}_{s,t}, \theta)$, which we will write simply as $\mathbb{P}(y|f(s,t), \theta)$, suppressing the spatio-temporal indices for ease of notation. Naturally, $f(s,t)$ $[= \mathbf{f}_{s,t}]$ is the realisation of a Gaussian process with mean function, $\mu_\theta(\cdot)$, and covariance function, $k_\theta(\cdot,\cdot)$, controlled by elements of a low-dimensional vector of hyperparameters, θ. Formally, the Gaussian process is defined as an (s,t)-indexed stochastic process for which the joint distribution over any finite collection of points, $(s,t)[i]$, is multivariate Gaussian with mean vector, $\mu_i = m((s,t)[i]|\theta)$, and covariance matrix, $\Sigma_{i,j} = k((s,t)[i],(s,t)[j]|\theta)$. The Bayesian hierarchy is completed by defining a vector of prior distributions for θ, which may potentially include hyperparameters for the likelihood (e.g., over-dispersion in a beta-binomial) in addition to those on parametrising the mean and covariance functions e.g the mean function coefficients β. In hierarchical notation, supposing for clarity an independent and identically distributed (iid) Normal likelihood with variance, σ_e^2:

$$\begin{aligned} \theta &\sim \pi(\theta) \\ f(s,t)|\mathbf{X}_{s,t}, \theta &\sim GP(\mu_\theta, k_\theta) \\ y|f(s,t), \mathbf{X}_{s,t}, \theta &\sim N(f(s,t), \mathbb{1}\sigma_e^2) \end{aligned} \tag{1}$$

Following Bayes theorem the posterior distribution resulting from this hierarchy becomes:

$$\mathbb{P}(\theta, f(s,t)|y) = \frac{\mathbb{P}(y|f(s,t),\theta)\mathbb{P}(f(s,t)|\theta)\mathbb{P}(\theta)}{\int\int \mathbb{P}(y|f(s,t),\theta)\{d\mathbb{P}(f(s,t)|\theta)\}\{d\mathbb{P}(\theta)\}}, \quad (2)$$

where the denominator in Equation 2 is the marginal likelihood, $\mathbb{P}(y)$.

Given the hierarchical structure in Equation 1 and the conditional properties of Gaussian distributions, the conditional predictive distribution for the mean of observations, z $[= \mathbf{z}_{s',t'}]$, at location-time pairs, $(s',t')[j]$, for a given θ is also Gaussian with form:

$$\begin{aligned} z|y,\theta &\sim N(\mu^*, \Sigma^*) \quad (3) \\ \mu^* &= \mu_{(s',t')|\theta} + \Sigma_{(s',t'),(s,t)|\theta}\Sigma^{-1}_{y|(s,t),\theta}\left(y - \mu_{(s,t)|\theta}\right) \\ \Sigma^* &= \Sigma_{(s',t')|\theta} - \Sigma_{(s',t'),(s,t)|\theta}\Sigma^{-1}_{y|(s,t),\theta}\Sigma_{(s,t),(s',t')|\theta} \quad (4) \end{aligned}$$

where $\Sigma_{y|(s,t),\theta} = \left(\Sigma_\theta + \mathbb{1}\sigma_e^2\right)$. For specific details on the parametrisation of Σ see the Appendix

When examining the conditional expectation in equation 4 and splitting the summation into terms $\mu_{(s',t')|\theta}$ and $\Sigma_{(s',t'),(s,t)|\theta}\Sigma^{-1}_{y|(s,t),\theta}\left(y - \mu_{(s,t)|\theta}\right)$, it is clear that the first specifies a global underlying mean while the second augments the residuals from that mean by the covariance function. Clearly, if the mean function fits the data perfectly the covariance in the second term of the expectation would drop out and conversely if the mean function is zero, then only the covariance function would model the data. This expectation therefore represents a balance between the underlying trend and the residual correlated noise.

In most applications of Gaussian process regression a linear mean function ($\mu_\theta = \mathbf{X}_{s,t}\beta$) is used, where β is a vector of m coefficients. However, when a rich suite of covariates is available this linear mean may be sub-optimal, limiting the generalisation accuracy of the model. To improve on the linear mean, covariate basis terms can be expanded to include parametric nonlinear transforms and interactions, but finding the optimal set of basis is computationally demanding and often leaves the researcher open to data snooping [38]. In this paper we propose using an alternative two stage statistical procedure to first obtain a set of candidate non-linear mean functions using multiple different algorithmic methods fit without reference to the assumed spatial covariance structure and then include those means in the Gaussian process via stacked generalisation.

Stacked generalisation

Stacked generalisation [36], also called stacked regression [37], is a general ensemble approach to combining different models. In brief, stacked generalisers combine different models together to produce a meta-model with equal or better predictive performance than the constituent parts [39]. In the context of malaria mapping our goal is to fuse multiple algorithmic methods with Gaussian process regression to both fully exploit the information contained in the covariates and model spatio-temporal correlations.

To present stacked generalisation we begin by introducing standard ensemble methods and show that stacked generalisation is simply a special case of this powerful technique. To simplify notation we suppress the spatio-temporal index and dependence on θ. Consider $\mathcal{L}$ models, with outputs $\tilde{y}_i(x), i = 1, ..., \mathcal{L}$. The choice of these models is described in the supplementary information. We denote the true target function as $f(x)$ and can therefore write the regression equation as $y_i(x) = f(x) + \epsilon_i(x)$. The average sum-of-squares error for model i is defined as $E_i = \mathbb{E}[(\tilde{y}_i(x) - f(x))^2]$. Our goal is to estimate an ensemble model across all $\mathcal{L}$ models, denoted as $M(\tilde{y}_1, .., \tilde{y}_\mathcal{L})$. The simplest choice for C is an average across all models $M(\tilde{y}_1, .., \tilde{y}_\mathcal{L}) = \tilde{y}_{\text{avg}}(x) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \tilde{y}_i(x)$. However this average assumes that the error of all models are the same, and that all models perform equally well. The assumption of equal performance may hold when using variants of a single model (i.e. bagging) but is unsuitable when very different models are used. Therefore a simple extension would be to use a weighted mean across models $M(\tilde{y}_1, .., \tilde{y}_\mathcal{L}) = \tilde{y}_{\text{wavg}}(x) = \sum_{i=1}^{\mathcal{L}} \beta_i \tilde{y}_i(x)$ subject to constraints $\beta > 0 \; \forall i, \sum_{i=1}^{\mathcal{L}} \beta = 1$ (convex combinations). These constraints prevent extreme predictions in well predicting models and impose the sensible inequality $\tilde{y}_{\min}(x) \leq \tilde{y}_{\text{wavg}}(x) \leq \tilde{y}_{\max}$ [37]. The optimal βs can be found by quadratic programming or by Bayesian linear regression with a Dirichlet/categorical prior on the coefficients. One particularly interesting result of combining models using this constrained weighted mean is the resulting decomposition of error into two terms [40]

$$\begin{aligned} \mathbb{E}[(\tilde{y}_{\text{wavg}}(x) - f(x))^2] = & \sum_{i=1}^{n} \beta_i \mathbb{E}[(\tilde{y}_i(x) - f(x))^2] \\ & - \sum_{i=1}^{n} \beta_i \mathbb{E}[(\tilde{y}_i(x) - \tilde{y}_{\text{wavg}}(x))^2] \end{aligned} \tag{5}$$

Equation 5 is a reformulation of the standard bias-variance decomposition [35] where first term describes the average error of all models and the second (termed the ambiguity) is the spread of each member of the ensemble around the weighted mean, measuring the disagreement among models. Equation 5 shows that combining multiple models with low error but with large disagreements produces a lower overall error. It should be noted that equation 5 makes the assumption that $y(x) = f(x)$.

Combination of models in an ensemble as described above can potentially lead to reductions in errors. However the ensemble models introduced so far are based only on training data and therefore neglect the issue of model complexity and tell us nothing about the ability to generalise to new data. To state this differently, the constrained weighted mean model will always allocate the highest weight to the model that most over fits the data. The standard method of addressing this issue is to use cross validation as a measure of the generalisation error and select the best performing of the $\mathcal{L}$ models. Stacked generalisation provides a technique to combine the power of ensembles described above but also produces models that can generalise well to new data. The principle idea behind stacked generalisation is to train $\mathcal{L}$ models (termed level 0 generalisers) and generalise their combined behaviour via a second model (termed the level 1 generaliser). Practically this is done by specifying a K-fold cross validation set, training all $\mathcal{L}$ level 0 models on these sets and using the cross validation predictions to train a level 1 generaliser. This calibrates the level 1 model based on the generalisation ability of the level 0 models. After this level 1 calibration, all level 0 models are refitted using the full data set and these predictions are used in the level 1 model without refitting (This procedure is more fully described in algorithm 1 and the schematic design shown in supplementary information). The combination of ensemble modelling with the ability to generalise well has made stacking one of the best methods to achieve state-of-the art predictive accuracy [37, 39, 41].

Defining the most appropriate level 1 generaliser based on a rigorous optimality criteria is still an open problem, with most applications using the constrained weighted mean specified above [37, 39]. Using the weighted average approach can be seen as a general case of cross validation, where standard cross validation would select a single model by specifying a single β_i as 1 and all other β_is as zero. Additionally, it has been shown that using the constrained weighted mean method will perform asymptotically as well as the best possible choice among

the family of weight combinations [39].

Here we suggest using Gaussian process regression as the level 1 generaliser. Revisiting equation 4 we can replace $\mu_{(s',t')|\theta}$ with a linear stacked function $\mu_{(s',t')|\theta} = \sum_{i=1}^{\mathcal{L}} \beta_i \tilde{y}_i(s',t')$ across $\mathcal{L}$ level 0 generalisers, where the subscript denotes predictions from the ith level 0 generaliser (see Algorithm 1. We also impose inequality constraints on β_i such that $\beta_i > 0 \forall i, \sum_{i=1}^{\mathcal{L}} \beta_i = 1$. This constraint allows the βs to approximately sum to one and helps computational tractability. It should be noted that empirical analysis suggests that simply imposing $\beta_i > 0 \forall i$, is practically sufficient [37].

The intuition in this extended approach is that the stacked mean of the Gaussian process uses multiple different methods to exploit as much predictive capacity from the covariates as possible and then leaves the spatio-temporal residuals to be captured through the Gaussian process covariance function. In the supplementary information we prove that this approach yields all the benefits of using constrained weighted mean (equation 5) but allows for a further reduction in overall error from the covariance function of the Gaussian process.

We note here that that stacked generalisers are distinct from Bayesian model averaging (BMA). Stacked generalisers expand and change the hypothesis space from which the learning algorithm chooses a function (e.g from single decision trees to a linear combination of them) and can take a variety of different forms. BMA, however, weights hypotheses from the original space according to a fixed formula [42]. Due to these fundamental differences previous studies have suggested suggested the stacking has better robustness properties than BMA in the most important settings [43].

Data, Covariates and Experimental Design

The hierarchical structure most commonly used in infectious disease mapping is that shown in Equation 1. In malaria studies our response data are discrete random variables representing the number of individuals testing positive for the *Plasmodium falciparum* malaria parasite, N^+, out of the total number tested, N, at a given location. If the response is aggregated from the individual household level to a cluster or enumeration area level, the centroid of the component

sites is used as the spatial position datum. The ratio of N^+ to N is defined as the parasite rate or prevalence and is a key epidemiological parameter measuring transmission intensity. The response data was additionally transformed via the empirical logit [1, 4]. Pre-modelling standardisation of the available prevalence data for age and diagnostic type has also been performed on the data used here, as described in depth in [4, 7]. Our analysis is performed over Sub-Saharan Africa with the study area and dataset partitioned into 4 epidemiologically-distinct regions [7]—Eastern Africa, Western Africa, North Eastern Africa, and Southern Africa—each of which was modelled separately (see Figure 1). The data using in this study is identical to that recently published by Bhatt et al 2015 [4] and collection process has been described in detail previously [4, 7, 8].

All the malaria response data are freely available through an online data explorer portal found at http://www.map.ox.ac.uk/. All the covariate grids are freely available and can be accessed at https://earthengine.google.com/datasets/. The code used in this analysis is freely available at https://codeshare.io/5wnRn7. Fitting and analysis was performed in the R programming language using the INLA, H2O, mgcv and earth packages. More information can be found in the supplementary information.

The covariates (i.e., independent variables) used in this research consist of raster layers spanning the entire continent at a 2.5 arc-minute (5 km x 5 km) spatial resolution. The majority of these raster covariates were derived from high temporal resolution satellite images that were first gap filled [44] to eliminate missing data (resulting primarily from persistent cloud cover over equatorial forests) and then aggregated to create a dynamic (i.e. temporally varying) dataset for every month throughout the study period (2000-2015). The list of covariates is presented in Table 1 and detailed information on individual covariates can be found here [24, 26, 44]. The set of monthly dynamic covariates was further expanded to include lagged versions of the covariate at 2 month, 4 month and 6 month lags. The main objective of this study was to judge the predictive performance of the various generalisation methods and therefore no variable selection or thinning of the covariate set was performed. It should be noted however that many of the level 0 generalisers performed variable selection automatically (e.g. elastic net regression).

The resolution used throughout was defined by the covariate grids at 5 km x 5 km. The

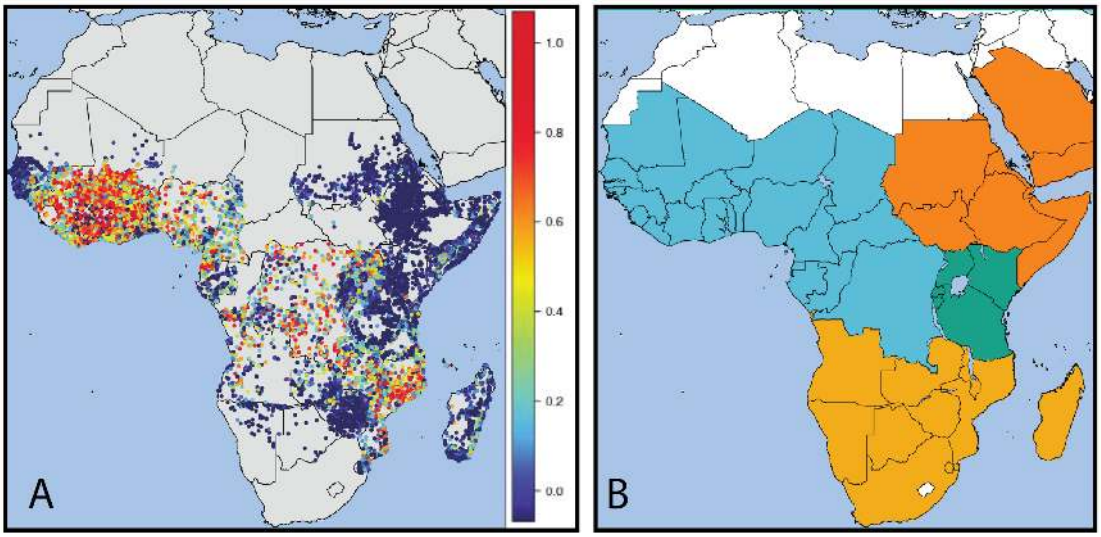
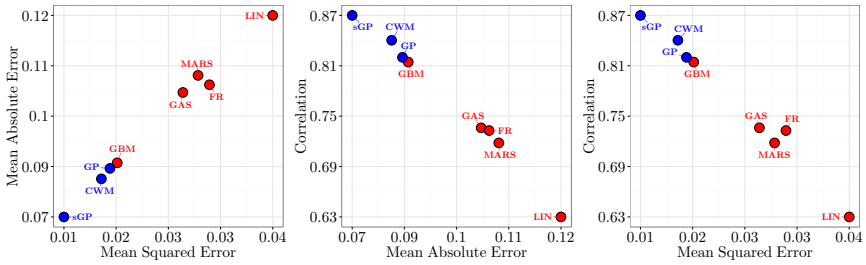


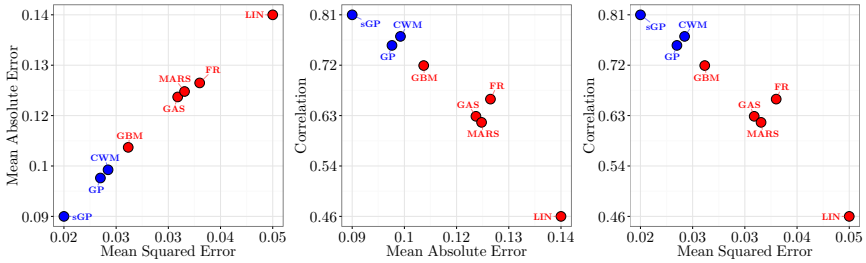
Figure 1. (A) Plot of the 23,131 prevalence surveys conducted between 2000 and 2015. The survey data are age and diagnostic standardized and presented as a continuum of blue to red from $0-1$ (B) Study area of stable malaria transmission in Sub-Saharan Africa. Our analysis was performed on 4 zones - Western Africa, North eastern Africa, Eastern Africa and Southern Africa

prevalence points were therefore snapped to the centroid of the pixel containing them. If multiple cluster points were contained within the same pixel at the same time, then they were aggregated. Likewise, the spatial field, which can be projected or evaluated at any spatial resolution, was taken as the value of the spatial field at the centroid of the pixel.

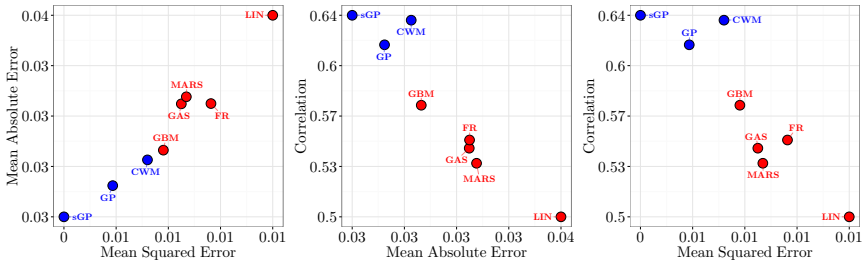
The level 0 generalisers used were gradient boosted trees [29, 45], random forests [30], elastic net regularised regression [34], generalised additive splines [27, 32] and multivariate adaptive regression splines [33]. The level 1 generalisers used were stacking using a constrained weighted mean and stacking using Gaussian process regression. We also fitted a standard Gaussian process for benchmark comparisons with the level 0 and 1 generalisers. Stacked fitting was performed following Algorithm 1. Full analysis and K-fold Cross validation was performed 5 times and then averaged to reduce any bias from the choices of cross validation set. The averaged cross validation results were used to estimate the generalisation error by calculating the mean squared error (MSE $(y - f)^2$)), mean absolute error (MAE$|y - f|$) and the correlation.



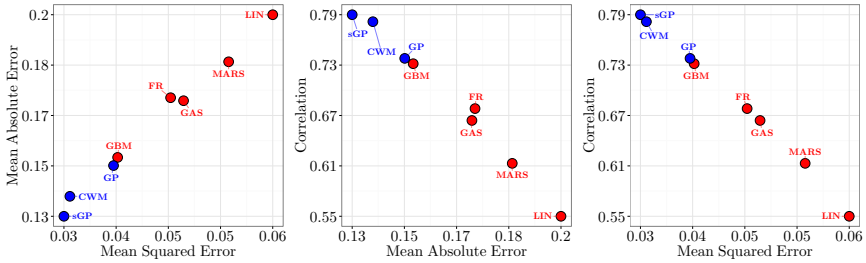
(a) Eastern Africa



(b) Southern Africa



(c) North Eastern Africa



(d) Western Africa

Figure 2. Comparisons of cross-validation MSE versus MAE versus correlation. Level 1 generalisers and the standard Gaussian process are shown in blue and all level 0 generalisers are shown in red. Legend abbreviations: (1) SGP - stacked Gaussian process, (2) CWM - stacked constrained weighted mean, (3) GP - standard Gaussian process, (4) GBM - Gradient boosted trees, (5) GAS - Generalised additive splines, (6) FR - Random forests, (7) MARS - Multivariate adaptive regression splines and (8) LIN - Elastic net regularised linear regression.

Results

The results of our analysis are summarised in Figure 2 where pairwise comparisons of MSE versus MAE versus correlation are shown. Across the Eastern, Southern and Western African regions (Figures 2a,2b and 2d), we found a consistent ranking pattern in the generalisation performance with the stacked Gaussian process approach presented in this paper outperforming all other methods. The constrained weighted mean stacked approach was the next best method followed by the standard Gaussian process (with a linear mean) and Gradient boosted trees. Random forests, multivariate adaptive regression splines and generalised additive splines all had similar performance and the worst performing method was the elastic net regularised regression. For the North Eastern region (Figure 2c), again the stacked Gaussian process approach was the best performing method but the standard Gaussian process performed better than the constrained weighted mean stacked approach, though only in terms of MAE and MSE.

One average, across all regions, the stacked Gaussian process approach reduced the MAE and MSE by 9% $[1\% - 13\%]$ (values in square brackets are the minimum and maximum across all regions) and 16% $[2\% - 24\%]$ respectively and increased the correlation by 3% $[1\% - 5\%]$ over the next best constrained weighted mean approach thereby empirically reinforcing the theoretical bounds derived in the supplementary information proof. When compared the the widely used elastic net linear regression the relative performance increase of the Gaussian process stacked approach is stark, with reduced MAE and MSE of 25% $[12\% - 33\%]$ and 25% $[19\% - 30\%]$ respectively and increase in correlation by 39% $[20\% - 50\%]$.

Compared to the standard Gaussian process previously used in malaria mapping the stacked Gaussian process approach reduced MAE and MSE by 10% $[3\% - 14\%]$ and 18% $[9\% - 26\%]$ respectively and increased the correlation by 6% $[3\% - 7\%]$.

Consistently across all regions the best non-stacked method was the standard Gaussian process with a linear mean function. Of the level 0 generalisers gradient boosted trees were the best performing method, with performance close to that of the standard Gaussian process. The standard Gaussian process only had a modest improvement over Gradient boosted trees with average reductions in MAE and MSE of 4% $[1\% - 8\%]$ and 7% $[1\% - 13\%]$ respectively and

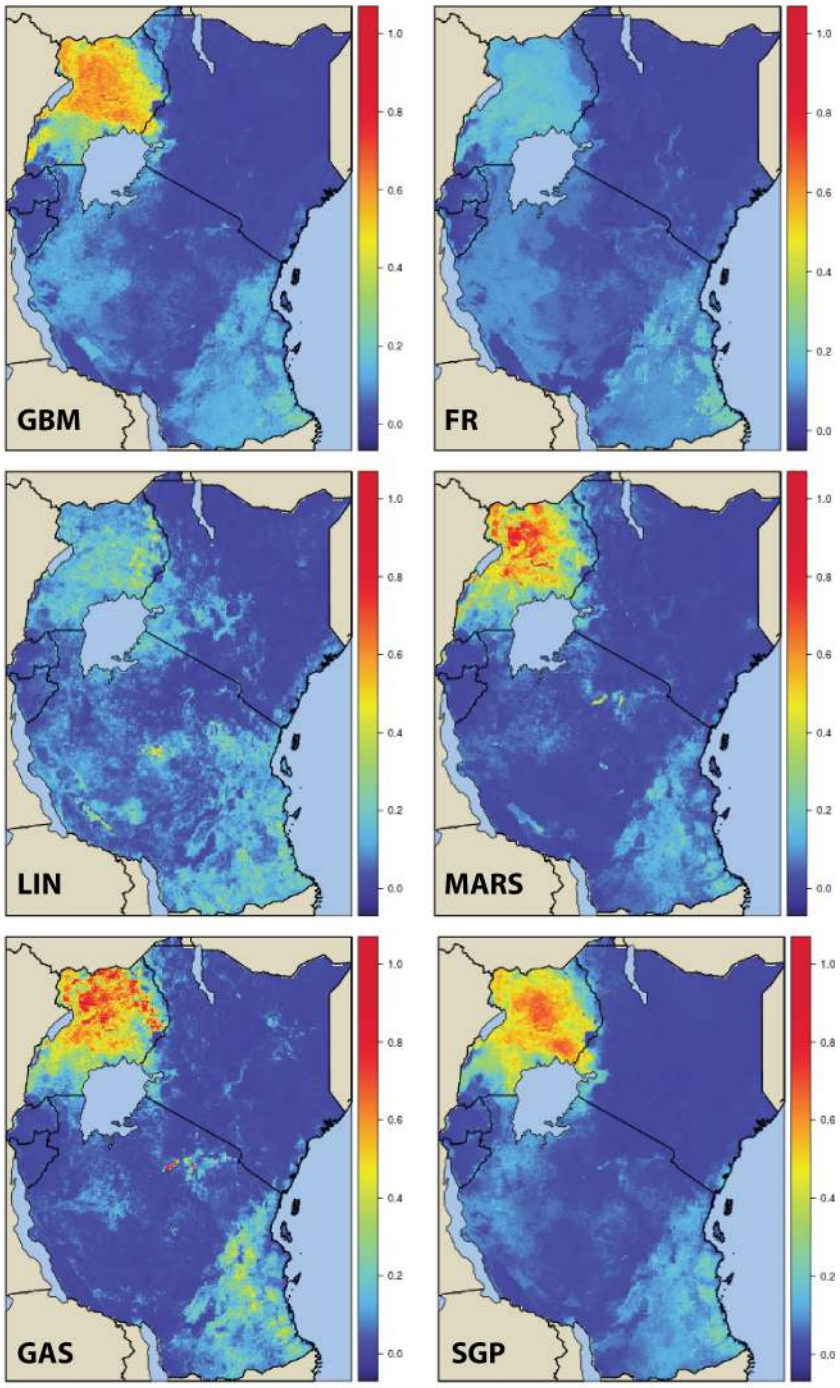


Figure 3. Predicted prevalence maps for Eastern Africa in 2011 for gradient boosted trees (GBM), random forests (FR), Elastic net regularised linear regression (LIN), Multivariate adaptive regression splines (MARS), generalised additive splines (GAS) and the new stacked Gaussian process (SGP)

increases in correlation of 3% [1% − 7%].

Figure 3 shows predicted map for all level 0 generalisers and the stacked Gaussian process approach for 2011 in the Eastern Africa region. There are clear similarities in the high and low regions across all maps and a strong correspondence to previous approaches [4, 7, 8]. The final ensemble map can be seen as a consensus of the the individual level 0 maps where the stacking algorithm weights each map according to generalisation performance. This is why the final stacked Gaussian process map most resembles the gradient boosted tree approach (the best predicting method, see Figure 2a) as opposed to the elastic net regularised linear regression approach (the worst predicting method). However, some idiosyncrasies of the gradient boosted approach, such as the sharp transition line in Southern Tanzania, are corrected in the stacked Gaussian process approach thanks to the other level 0 methods and the addition of spatio-temporal correlation.

Discussion

All the level 0 generalisation methods used in this paper have been previously applied to a diverse set of machine learning problems and have track records of good generalisability [27]. For example, in closely related ecological applications, these level 0 methods have been shown to far surpass classical learning approaches [46]. However, as introduced by Wolpert [36], rather than picking one level 0 method, an ensemble via a second generaliser has the ability to improve prediction beyond that achievable by the constituent parts [40]. Indeed, in all previous applications [36, 37, 39, 47] ensembling by stacking has consistently produced the best predictive models across a wide range of regression and classification techniques. The most popular level 1 generaliser is the constrained weighted mean with convex combinations. The key attraction of this level 1 generaliser is the ease of implementation and theoretical properties [39, 40]. In this paper we show that, for disease mapping, stacking using Gaussian processes is more predictive and generalises better than both single level 0 generalisers in isolation, and the more common stacking approach using a constrained weighted mean.

The key benefit of stacking is summarised in equation 5 where the total error of an ensemble

model can be reduced by using multiple, very different, but highly predictive models. However, stacking using a constrained weighted mean only ensures that the predictive power of the covariates are fully utilised and does not exploit the predictive power that could be gained from characterising any residual covariance structure. The standard Gaussian process suffers from the inverse situation where the covariates are underexploited and predictive power is instead gained from leveraging residual spatio-temporal covariance. In a standard Gaussian process the mean function is usually paramaterised through simple linear basis functions [48] that are often unable to model the complex non linear interactions needed to correctly capture the true underlying mean. This inadequacy is best highlighted by the poor generalisation performance of the elastic net regularised regression method across all regions. The trade off between the variance explained by the covariates versus that explained by the covariance function will undoubtedly vary from setting to setting. For example in the Eastern, Southern, and Western African regions, the constrained weighted mean stacking approach performs better than the standard Gaussian process and the level 0 gradient boosted trees generaliser performs almost as well as the standard Gaussian process. For these regions, this shows a strong influence of the covariates on the underlying process. In contrast, for the North Eastern African region, the standard Gaussian process does better than both the constrained weighted mean approach (in terms of error not correlation) and all of the level 0 generalisers, suggesting a weak influence of the covariates. However, for all zones, the stacked Gaussian process approach is consistently the best approach across all predictive metrics. By combining both the power of Gaussian processes to characterise a complex covariance structure, and multiple algorithmic approaches to fully exploit the covariates, the stacked Gaussian process approach combines the best of both worlds and predicts well in all settings.

This paper introduces one way of stacking that is tailored for spatio-temporal data. However the same principles are applicable to purely spatial or purely temporal data, settings in which Gaussian process models excel. Additionally, there is no constraint on the types of level 0 generalisers than can be used; dynamical models of disease transmission e.g. Malaria mechanistic models [49] [50] can be fitted to data and used as the mean function within the stacked framework. Using dynamical models in this way can constrain the mean to include known biological mechanisms that can potentially improve generalisability, allow for forecast predictions,

and help restrict the model to only plausible functions when data is sparse. Finally multiple different stacking schemes can be designed (see the supplementary information for details) and relaxations on linear combinations can be implemented (e.g. [47]).

Gaussian processes are increasingly being used for expensive optimisation problems [51] and Bayesian quadrature [52]. In current implementations both of these applications are limited to low dimensional problems typically with less than 10 parameters. Future work will explore the potential for stacking to extend these approaches to high dimensional settings. The intuition is that the level 0 generalisers can accurately and automatically learn much of the latent structure in the data, including complex features like non-stationarity, which are a challenge for Gaussian processes. Learning this underlying structure through the mean can leave a much simpler residual structure [25] to be modelled by the level 1 Gaussian process.

In this paper we have focused primarily on prediction, that is neglecting any causal inference and only searching for models with the lowest generalisation error. Determining causality from the complex relationships fitted through the stacked algorithmic approaches is difficult but empirical methods such as partial dependence [29] or individual conditional expectation [53] plots can be used to approximate the marginal relationships from the various covariates. Similar statistical techniques can also be used to determine covariate importance.

Increasing volumes of data and computational capacity afford unprecedented opportunities to scale up infectious disease mapping for public health uses [54]. Maps of diseases and socio-economic indicators are increasingly being used to inform policy [4, 55], creating demand for methods to produce accurate estimates at high spatial resolutions. Many of these maps can subsequently be used in other models but, in the first instance, creating these maps requires continuous covariates, the bulk of which come from remotely sensed sources. For many indicators, such as HIV or Tuberculosis, these remotely sensed covariates serve as proxies for complex phenomenon and as such, the simple mean functions in standard Gaussian processes are insufficient to predict with accuracy and low generalisation error. The stacked Gaussian process approach introduced in this paper provides an intuitive, easy to implement method that predicts accurately through exploiting information in both the covariates and covariance structure.

Algorithm 1 Stacked Generalisation Algorithm: The algorithm proceeds as follows. In line 2 to 4 the covariates,response and number of cross validation folds is defined. Lines 6 to 9 fits all level 0 generalisers to the full data set. Lines 10 to 16 fits all level 0 generalisers to cross validation data sets. Line 17 to 18 fits a level 1 generaliser to the cross validation predictions and Line 19 returns the final output by using the level 1 generaliser to predict on the full predictions

1: **procedure** STACK ▷ covariate and response input
2: **Input** X as a $n \times m$ design matrix
3: **Input** y as a n vector of responses
4: **Input** v cross validation folds
5: **choose** l, $\mathcal{L}(y, X)$ models ▷ level 0 generalisers
6: **define** $n \times l$ matrix P ▷ matrix of predictions
7: **for** $i \leftarrow 1, l$ **do**
8: **fit** $\mathcal{L}_i(y, X)$
9: **predict** $P_{,i} = \mathcal{L}_i(y, X)$
10: **split** X, y into $\{g_1, ...g_v\}$ groups $\{X_{g_1}, .., X_{g_v}\}$ and $\{y_{g_1}, .., y_{g_v}\}$ ▷ training set
11: **add** remaining samples to $\{X_{/g_1}, .., X_{/g_v}\}$ and $\{y_{/g_1}, .., y_{/g_v}\}$ ▷ testing set
12: **define** $n \times l$ matrix H ▷ matrix cross validation of predictions
13: **for** $i \leftarrow 1, l$ **do**
14: **for** $j \leftarrow 1, v$ **do**
15: **fit** $\mathcal{L}_i(y_{g_j}, X_{g_j})$
16: **predict** $H_{/g_j,i} = \mathcal{L}_i(y_{/g_j}, X_{/g_j})$
17: **choose** $\mathcal{L}^*(y, H)$ model ▷ level 1 generaliser
18: **fit** $\mathcal{L}^*(y, H)$
19: **return** $\mathcal{L}^*(y, P)$ ▷ final prediction output

Table 1. List of Environmental, Socio-demographic and Land type covariates used.

Variable Class	Variable(s)	Source	Type
Temperature	Land Surface Temperature (day, night, and diurnal-flux)	MODIS Product	Dynamic Monthly
Temperature Suitability	Temperature Suitability for Plasmodium falciparum	Modeled Product	Dynamic Monthly
Precipitation	Mean Annual Precipitation	WorldClim	Synoptic
Vegetation Vigor	Enhanced Vegetation Index	MODIS Derivative	Dynamic Monthly
Surface Wetness	Tasseled Cap Wetness	MODIS Derivative	Dynamic Monthly
Surface Brightness	Tasseled Cap Brightness	MODIS Derivative	Dynamic Monthly
IGBP Landcover	Fractional Landcover	MODIS Product	Dynamic Annual
IGBP Landcover Pattern	Landcover Patterns	MODIS Derivative	Dynamic Annual
Terrain Steepness	SRTM Derivatives	MODIS Product	Static
Flow & Topographic Wetness	Topographically Redistributed Water	SRTM Derivatives	Static
Elevation	Digital Elevation Model	SRTM	Static
Human Population	AfriPop	Modeled Products	Dynamic Annual
Infrastructural Development	Accessibility to Urban Centers and Nighttime Lights	Modeled Product and VIIRS	Static
Moisture Metrics	Aridity and Potential Evapotranspiration	Modeled Products	Synoptic

Author Contributions

Conceived of and designed the research: SB. Drafted the manuscript: SB and EC. Drafted the supplementary information: SB. Prepared data: DJW. Conducted the analyses: SB. Supported the analyses: SB, EC, SRF. Supported interpretation and policy contextualization: DLS and PWG. All authors discussed the results and contributed to the revision of the final manuscript.

Funding Statement

SB is supported by the MRC outbreak centre and the Bill and Melinda Gates Foundation [*opp*1152978]. DLS was supported by the Bill and Melinda Gates Foundation [$OPP1110495$], National Institutes of Health/National Institute of Allergy and Infectious Diseases [$U19AI089674$], and the Research and Policy for Infectious Disease Dynamics (RAPIDD) program of the Science and Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. PWG is a Career Development Fellow [$K00669X$] jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement, also part of the EDCTP2 programme supported by the European Union, and receives support from the Bill and Melinda Gates Foundation [$OPP1068048, OPP1106023$]. These grants also support DJW, and EC.

Acknowledgements

We would like to Acknowledge Mike Thorne for proof reading the manuscript.

Data Accessibility

All the data used in this paper is feely accessible. All the malaria response data are freely available through an online data explorer portal found at http://www.map.ox.ac.uk/. All the covariate

grids are freely available and can be accessed at https://earthengine.google.com/datasets/. The code used in this analysis is freely available at https://codeshare.io/5wnRn7. Fitting and analysis was performed in the R programming language using the INLA, H2O, mgcv and earth packages. More information can be found in the supplementary information.

Ethics

Not applicable

References

1. Peter Diggle and PJ Ribeiro. *Model-based Geostatistics*. Springer, New York, 2007.

2. S.I. Hay, K.E. Battle, D.M. Pigott, D.L. Smith, C.L. Moyes, S. Bhatt, J.S. Brownstein, N. Collier, M.F. Myers, D.B. George, and P.W. Gething. Global mapping of infectious disease. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614), 2013.

3. Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association : JAMIA*, 15(2):150–7, 1 2008.

4. S Bhatt, D J J Weiss, E Cameron, D Bisanzio, B Mappin, U Dalrymple, K E E Battle, C L L Moyes, A Henry, P A A Eckhoff, E A A Wenger, O Briët, M A A Penny, T A A Smith, A Bennett, J Yukich, T P P Eisele, J T T Griffin, C A A Fergus, M Lynch, F Lindgren, J M M Cohen, C L J L J Murray, D L L Smith, S I I Hay, R E E Cibulskis, and P W W Gething. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526(7572):207–211, 9 2015.

5. C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*, volume 14. 2006.

6. Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. 2006.

7. Peter W Gething, Anand P Patil, David L Smith, Carlos a Guerra, Iqbal R.F. Elyazar, Geoffrey L Johnston, Andrew J Tatem, and Simon I Hay. A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malaria journal*, 10(1):378, 1 2011.

8. Simon I Hay, Carlos a Guerra, Peter W Gething, Anand P Patil, Andrew J Tatem, Abdisalan M Noor, Caroline W Kabaria, Bui H Manh, Iqbal R F Elyazar, Simon J Brooker, David L Smith, Rana a Moyeed, Robert W Snow, Kabaria C.W., Bui H Manh, Iqbal R F Elyazar, Simon J Brooker, David L Smith, Rana a Moyeed, and Robert W Snow. A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med*, 6(3):0286–0302, 3 2009.

9. Laura Gosoniu, Amina Msengwa, Christian Lengeler, and Penelope Vounatsou. Spatially explicit burden estimates of malaria in Tanzania: bayesian geostatistical modeling of the malaria indicator survey data. *PloS one*, 7(5):e23966, 1 2012.

10. Abbas B Adigun, Efron N Gajere, Olusola Oresanya, and Penelope Vounatsou. Malaria risk in Nigeria: Bayesian geostatistical modelling of 2010 malaria indicator survey data. *Malaria journal*, 14(1):156, 1 2015.

11. Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

12. Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

13. H Rue, S Martino, and N Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2009.

14. Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in medicine*, 29(15):1580–1607, 2010.

15. Thomas P. Minka. Expectation propagation for approximate Bayesian inference. pages 362–369, 8 2001.

16. James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

17. Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21(3):786–92, 3 2009.

18. Somak Dutta and Debashis Mondal. REML estimation with intrinsic Matérn dependence in the spatial linear mixed model. *Electronic Journal of Statistics*, 10(2):2856–2893, 2016.

19. A Rahimi and B Recht. Random features for large-scale kernel machines. *Advances in neural information processing*, 2008.

20. Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

21. Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

22. David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013.

23. Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian process kernels for pattern discovery and extrapolation. *arXiv preprint arXiv:1302.4245*, 2013.

24. Daniel J J Weiss, Bonnie Mappin, Ursula Dalrymple, Samir Bhatt, Ewan Cameron, Simon I I Hay, and Peter W W Gething. Re-examining environmental correlates of Plasmodium

falciparum malaria endemicity: a data-intensive variable selection approach. *Malaria journal*, 14(1):68, 1 2015.

25. Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14:505–531, 10 2015.

26. Daniel J J Weiss, Samir Bhatt, Bonnie Mappin, Thomas P P Van Boeckel, David L L Smith, Simon I I Hay, and Peter W W Gething. Air temperature suitability for Plasmodium falciparum malaria transmission in Africa 2000-2012: a high-resolution spatiotemporal prediction. *Malaria journal*, 13(1):171, 1 2014.

27. Trevor Hastie, Robert Tibshirani, and J H Friedman. *The elements of statistical learning*. Springer, 2009.

28. Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

29. J H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

30. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

31. Robert Tibshirani Trevor Hastie. Generalized Additive Models. *Statistical Science*, 1(3):297–310, 1986.

32. Simon Wood. *Generalized Additive Models: An Introduction with R*. CRC Press, 2006.

33. Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, 3 1991.

34. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 4 2005.

35. S Geman, E Bienenstock, and R Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 1992.

36. David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1 1992.

37. Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

38. YS Abu-Mostafa, M Magdon-Ismail, and HT Lin. *Learning from data*. 2012.

39. Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article25, 1 2007.

40. Anders Krogh and Jesper Vedelsby. Neural Network Ensembles, Cross Validation, and Active Learning. In *Advances in Neural Information Processing Systems*, pages 231–238. MIT Press, 1995.

41. Antti Puurula, Jesse Read, and Albert Bifet. Kaggle LSHTC4 Winning Solution. 5 2014.

42. Pedro Domingos and Pedro. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78, 10 2012.

43. Bertrand Clarke and Bertrand@stat Ubc Ca. Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored. *Journal of Machine Learning Research*, 4:683–712, 2003.

44. Daniel J Weiss, Peter M Atkinson, Samir Bhatt, Bonnie Mappin, Simon I Hay, and Peter W Gething. An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS journal of photogrammetry and remote sensing : official publication of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, 98:106–118, 12 2014.

45. J H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

46. J Elith, J R Leathwick, and T Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

47. Joseph Sill, Gabor Takacs, Lester Mackey, and David Lin. Feature-Weighted Linear Stacking. 11 2009.

48. C Rasmussen. Gaussian processes in machine learning. *Advanced Lectures on Machine Learning*, pages 63–71, 2004.

49. David L Smith and F Ellis McKenzie. Statics and dynamics of malaria infection in Anopheles mosquitoes. *Malaria journal*, 3(1):13, 6 2004.

50. Jamie T Griffin, Neil M Ferguson, and Azra C Ghani. Estimates of the changing age-burden of Plasmodium falciparum malaria disease in sub-Saharan Africa. *Nature communications*, 5, 2014.

51. A. O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 11 1991.

52. Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2179), 2015.

53. Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 3 2015.

54. David M Pigott, Rosalind E Howes, Antoinette Wiebe, Katherine E Battle, Nick Golding, Peter W Gething, Scott F Dowell, Tamer H Farag, Andres J Garcia, Ann M Kimball, L Kendall Krause, Craig H Smith, Simon J Brooker, Hmwe H Kyu, Theo Vos, Christopher J L Murray, Catherine L Moyes, and Simon I Hay. Prioritising Infectious Disease Mapping. *PLoS neglected tropical diseases*, 9(6):e0003756, 6 2015.

55. Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, Monica F Myers, Dylan B George, Thomas Jaenisch, G R William Wint, Cameron P Simmons, Thomas W Scott, Jeremy J Farrar, and Simon I Hay. The global distribution and burden of dengue. *Nature*, 496(7446):504–7, 4 2013.