Malaria Journal

RESEARCH

Open Access

Comparison of new computational methods for spatial modelling of malaria

Spencer Wong[1], Jennifer A. Flegg[1*], Nick Golding[2,3] and Sevvandi Kandanaarachchi[4]

Abstract

Background Geostatistical analysis of health data is increasingly used to model spatial variation in malaria prevalence, burden, and other metrics. Traditional inference methods for geostatistical modelling are notoriously computationally intensive, motivating the development of newer, approximate methods for geostatistical analysis or, more broadly, computational modelling of spatial processes. The appeal of faster methods is particularly great as the size of the region and number of spatial locations being modelled increases.

Methods This work presents an applied comparison of four proposed 'fast' computational methods for spatial modelling and the software provided to implement them—Integrated Nested Laplace Approximation (INLA), tree boosting with Gaussian processes and mixed effect models (GPBoost), Fixed Rank Kriging (FRK) and Spatial Random Forests (SpRF). The four methods are illustrated by estimating malaria prevalence on two different spatial scales—country and continent. The performance of the four methods is compared on these data in terms of accuracy, computation time, and ease of implementation.

Results Two of these methods—SpRF and GPBoost—do not scale well as the data size increases, and so are likely to be infeasible for larger-scale analysis problems. The two remaining methods—INLA and FRK—do scale well computationally, however the resulting model fits are very sensitive to the user's modelling assumptions and parameter choices. The binomial observation distribution commonly used for disease prevalence mapping with INLA fails to account for small-scale overdispersion present in the malaria prevalence data, which can lead to poor predictions. Selection of an appropriate alternative such as the Beta-binomial distribution is required to produce a reliable model fit. The small-scale random effect term in FRK overcomes this pitfall, but FRK model estimates are very reliant on providing a sufficient number and appropriate configuration of basis functions. Unfortunately the computation time for FRK increases rapidly with increasing basis resolution.

Conclusions INLA and FRK both enable scalable geostatistical modelling of malaria prevalence data. However care must be taken when using both methods to assess the fit of the model to data and plausibility of predictions, in order to select appropriate model assumptions and parameters.

Keywords Spatial modelling, Geostatistics, Predictive modelling, Risk mapping

*Correspondence:
Jennifer A. Flegg
jennifer.flegg@unimelb.edu.au
[1] School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia
[2] Telethon Kids Institute, Perth Children's Hospital, 15 Hospital Ave, Nedlands, WA 6009, Australia
[3] Curtin University, Kent St, Bentley, WA 6102, Australia
[4] CSIRO's Data61, Research Way, Clayton, VIC 3168, Australia

Background

Spatial proximity often plays an important role in governing the spread of geographic processes. Geostatistical techniques directly model these effects of proximity and are used to create continuous predictions from a finite set of observations. Since their original development for use in the mining sector [1], these techniques are now applied to wide ranging problems

where spatial correlation must be accounted for, including species distribution modelling in ecology [2], interpolating weather and climate data [3, 4], mapping soil properties for agriculture [5], and spatial modelling of real estate prices [6]. Mapping disease risk is an important application in epidemiology, where geostatistical techniques are used to predict risks for a wide range of diseases with differing biology [7–10]. These methods are particularly prominent in malaria work, where risk maps inform policy and elimination strategies [11, 12]. Predictive maps created using geostatistical models have been published for malaria prevalence [13–16], mortality [17], use of malaria interventions [18], antimalarial drug resistance [19, 20], risks of adverse effects [21], and the relationship between sickle cell anaemia and *P. falciparum* [22]. The mapped metrics depend on various spatial processes including environmental factors (e.g., rainfall and temperature), variable access to health care, and human movement. In the absence of a full understanding of all these processes, spatial statistical modelling aims to describe the spatial variation in the metric of interest that is caused by the underlying spatial processes.

In their book, Diggle and Ribeiro Jr. introduce a fundamental paradigm for modelling geospatial data that unites previous spatial modelling approaches with model-based statistical analysis [1]. The quantity of interest or the response, y_i, is defined throughout a contiguous study region and each measurement at the sample location x_i is a realization of the random variable Y_i whose distribution is dependent on the location x_i as well as the random variables associated with the other data locations. That is, the random variables in space are dependent on each other based on their proximity. Hence, the observed responses at n locations are modelled as a joint n-dimensional vector of random variables where the dependency can be modelled using spatial random effects as part of a generalized linear geostatistical model. The spatially-correlated random variables are modelled as a Gaussian process (GP) and spatial covariates, such as bio-climatic and environmental layers, are often included as additional regressors to capture general trends.

Gaussian processes are widely used in spatio-temporal modelling including in malaria prevalence mapping research [23–25]. With the explosion of machine learning research, the popularity of GPs has remarkably increased in both theoretical and applied domains [26]. Rasmussen and Nickisch made available a toolbox called *GPML* for machine learning regression and classification tasks [27]. GPs for large scale regression [28] and GPs for sparse approximations [29] are examples of the use of GPs in machine learning. These new advances have made GPs a viable tool for modelling of very large datasets beyond the field of malaria mapping [30].

These modelling approaches vary in their inference procedures (e.g. Bayesian or frequentist) and computational techniques (e.g. simulation versus optimization). One thing these newer methods have in common is that they have the potential to avoid calculations using a 'full' (approximation-free) GP, due to the fact that full GP models scale cubicly with the number of unique locations in the training data. That is, a tenfold increase in the number of unique spatial locations in the dataset results in a 1000-fold increase in computation time. Consequently, the full GP can become computationally infeasible for large datasets, such as those used in national- and continental-scale malaria mapping. For such large datasets, it may also be prohibitive to fit the model using asymptotically exact Bayesian methods such as Markov chain Monte Carlo (MCMC) methods, so deterministic approximations to such simulation approaches have also been explored. Newer approximation methods to both the spatial random effect and the inference method are often used to combat this limitation for example when using global-scale datasets [31, 32].

There are a multitude of approximation techniques available as alternatives to full Bayesian/frequentist inference, and the full Gaussian process, in addition to modern machine learning methods which lie outside of the standard geostatistical framework while still enabling the fitting of spatially-explicit models as used in disease mapping. Geostatistical techniques using Gaussian processes to model spatial autocorrelation are currently the most popular method for malaria risk mapping [33],with full Gaussian processes with both maximum likelihood frameworks for inference [14, 21] and Bayesian inference using MCMC [34, 35] appearing frequently in the literature. The Integrated Nested Laplace Approximation has become a well established method in the field [15, 36, 37], while recent modelling techniques outside of the popular geostatistical framework such as boosted regression trees and random forests have found some limited use [38, 39].

A review of all such alternative techniques is beyond the scope of this paper. Due to the importance of spatial modelling in the malaria mapping field and consequent need for computationally efficient methods however, this work presents a comparison of four such methods on a malaria prevalence mapping problem: Integrated Nested Laplace Approximation inference, with a Gaussian Markov Random Field approximation to the GP (INLA), Gaussian processes fitted via a boosting algorithm (GPBoost), Spatial Random Forests (SpRF), and Fixed Rank Kriging (FRK). These four methods are selected due

to their different underlying techniques for modelling the spatial correlation structure and varying approaches to inference. Of these four methods, only INLA has been applied to malaria risk mapping, while the other three methods have found use in spatial modelling applications in various other fields (see, for example [40–42]).

The intended audience of this comparison are twofold. This work may be of interest to researchers interested in moving into the spatial (or spatio-temporal) mapping field who are looking for an introduction to currently available methods, and it may additionally inform malaria mapping researchers on the relative strengths and weaknesses of the considered methods when applied to a malaria mapping problem, assisting in decision making for future work. It should be noted that the inclusion of a method in the analysis does not constitute an endorsement of its use for malaria risk mapping. Rather, the purpose of this work is to compare each of these method's suitability, strengths, and weaknesses, when applied to malaria risk mapping problems, and in particular to examine the viability of the newer and previously unused methods.

An analysis at a national scale is first presented, with Kenya selected as the country of interest. This is then extended to a continental-scale analysis over Africa. The four methods are briefly introduced in "Methods" section, where the models implemented using each method are additionally specified. Due to their underlying mathematical differences in model specification and inference procedure, it is difficult to directly compare results. This problem is mitigated by comparing point and interval predictions against observed data in a cross-validation scheme. In addition, predictive spatial maps produced by each of the implemented models are explored. National and continent scale results are discussed in "Case study: Kenya" and "Continent scale results" section, and the computation time taken by each of the methods at each scale is compared in "Computational results" section. As concluding remarks, nuances of the methods uncovered by the analysis are briefly discussed in "Discussion" section. The programming scripts for this work are available at [43].

Methods

Diggle and Ribeiro Jr. first introduce a basic geostatistical model that does not have any covariates [1]. They consider data given by $(\mathbf{x}_i, y_i)$ for $i \in \{1, \ldots, n\}$, where $\mathbf{x}_i$ denotes the spatial location (i.e. coordinates) and y_i is the measured value for the quantity of interest at that location (e.g. the incidence of malaria at $\mathbf{x}_i$). They describe a model for normally-distributed response data with a stationary Gaussian process (one that tends back to the same average value, over the whole analysis region) as:

$$\left\{S(\mathbf{x}) : \mathbf{x} \in \mathrm{R}^2\right\}$$

where $S(\mathbf{x})$ is a Gaussian process with mean $\boldsymbol{\mu}$ (the average value over the study region), variance or amplitude of the process at each location $\boldsymbol{\sigma}^2 = \mathrm{var}\{S(\mathbf{x})\}$ and correlation function $\boldsymbol{\rho}(u) = \mathrm{cor}\{S(\mathbf{x}), S(\mathbf{x}')\}$, where $u = ||\mathbf{x} - \mathbf{x}'||$ and $||\cdot||$ denotes Euclidean distance (which controls the similarity of responses based on their distances apart); and y_i are realizations of mutually independent Gaussian random variables Y_i conditional on $\{S(\mathbf{x}) : \mathbf{x} \in R^2\}$ (i.e. after accounting for the spatial correlation, each y_i is independent and normally-distributed).

The model can be described by the equation:

$$Y_i \sim N\left(z_i, \boldsymbol{\tau}^2\right)$$

$$z_i = S(\mathbf{x}_i) \text{ for } i = 1, \ldots, n. \tag{1}$$

This basic model represents only the effects of proximity on observations at different locations, and can be extended through the addition of a *mean function* to model the effects of covariates on the response. Common choices of correlation functions (termed covariance functions when they incorporate the variance term $\boldsymbol{\sigma}^2$) include Matérn, exponential and squared exponential functions.

INLA, GPBoost, and FRK provide approximation and inference tools for geostatistical models which extend the basic model in Eq. (1) by [1]. In contrast, SpRF avoids the use of a Gaussian process entirely, instead using a machine learning approach to model the impacts of spatial proximity on the response. Despite their varying approaches, all four methods allow for prediction and mapping of spatial processes, which is the focus in this paper.

The following implementations of the four methods are used:

1. INLA: Integrated Nested Laplace Approximations, implemented in the R package `INLA` [44, 45].
2. GPBoost: Tree boosting with Gaussian processes and mixed effect models, implemented in the R package `gpboost` [46].
3. SpRF: Spatial Random Forests, implemented in the R package `ranger` [47].
4. FRK: Fixed Rank Kriging, implemented in the R package `FRK` [48].

As the national-scale dataset, *P. falciparum* prevalence data in Kenya from 2009 is used, retrieved from the open-access portion of the Malaria Atlas Project malaria prevalence dataset. Kenya was selected as it had the highest number of surveys overall, with the most surveys occurring in 2009. In expanding to a continental scale, available surveys across Africa in 2009 are used, keeping the same year between analyses. The R package `malariaAtlas` [32] was used in order to download the malaria prevalence survey data. As the aim of this work is to evaluate the performance of statistical models for malaria mapping rather than to produce reliable maps per say, additional validation, correction, or selection on these datapoints were not applied. For each record, only the spatial coordinates, the numbers of individuals screened, and the number of those individuals that were positive for *P. falciparum* were extracted.

The models constructed in this paper represent the simplest possible implementations of each of the four methods, where factors such as environmental covariates are not included. Indeed, due to the strong role environmental factors play in the malaria parasite's and vectors' life cycles [49], an in-depth selection and analysis of covariates is a key step in creating useful maps of malaria [33]. In skipping this step, the goal of this work is to highlight the differences between the methods at their most basic level, and specifically their differing treatments of spatial autocorrelation. Hence, the risk maps presented in this work should be considered only for the purpose of comparison with one another, and should not be interpreted as realistic maps of malaria prevalence.

INLA

INLA (Integrated Nested Laplace Approximations) is a method for approximate Bayesian inference which offers an improvement in speed over asymptotically exact methods such as MCMC. Instead of estimating a high-dimensional joint posterior distribution by simulation, INLA obtains approximations to univariate posterior marginal distributions of the model parameters. INLA is restricted to the class of models that can be expressed as latent Gaussian Markov random fields. However, a multitude of commonly used models can be expressed in this form, including generalized linear geostatistical models. This approach to inference pairs well with an approximation to the spatial Gaussian process as a Gaussian Markov random field (GMRF) over a discrete 'mesh' describing the study area, with piecewise linear interpolation to any locations that fall between nodes of this 'mesh'. When the Gaussian process has a covariance function of the Matérn type, the stochastic partial differential equation (SPDE) representation of the GMRF can be used, which makes evaluation of the spatial process very fast for large spatial datasets, compared with the full GP approach. Over the years there have been many updates to INLA [50] to broaden its scope and facilitate diverse problem solving tasks. For more details, refer to their website [51].

Inference with INLA combines a series of assumptions and Laplace approximations to compute the marginal posteriors of model parameters and latent effects. INLA assumes that the response vector $\mathbf{y}$ depends on a vector of latent variables $\boldsymbol{\eta}$, and hyperparameters $\boldsymbol{\theta}_1$, with density $\pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta}_1)$. The latent variables for example may include the values of a linear predictor, an intercept, regression coefficients, and the values of any random effects. Importantly, $\boldsymbol{\eta}$ is assumed to be a mean 0 Gaussian Markov random field with precision matrix $\mathbf{Q}(\boldsymbol{\theta}_2)$ (the construction of $\mathbf{Q}$ for continuous spatial models is outlined in [52]) where $\boldsymbol{\theta}_2$ is a vector of hyperparameters. The hyperparameters are often combined into a single vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with prior distribution $\pi(\boldsymbol{\theta})$. INLA then approximates the marginal posteriors $\pi(\boldsymbol{\eta}_i|\mathbf{y})$ and $\pi(\boldsymbol{\theta}_k|\mathbf{y})$ as follows.

The first step is to write the joint posterior of the hyperparameters as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})}$$

$$\propto \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})}{\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})}.$$

A Laplace approximation is applied to the denominator, replacing it with a Gaussian and giving the approximation:

$$\widetilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})}{\widetilde{\pi_G}(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\boldsymbol{\eta}=\eta^*(\boldsymbol{\theta})}, \tag{2}$$

where $\eta^*(\boldsymbol{\theta})$ is the mode of $\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})$, and $\widetilde{\pi_G}(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})$ is its Gaussian approximation. Approximate posterior marginals for the hyperparameters can then be obtained as:

$$\widetilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y}) = \int \widetilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-k}.$$

The exact marginals for the latent effects:

$$\pi(\boldsymbol{\eta}_i|\mathbf{y}) = \int \pi(\boldsymbol{\eta}_i|\boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

are approximated using numerical integration as

$$\widetilde{\pi}(\boldsymbol{\eta}_i|\mathbf{y}) = \sum_{k=1}^{K} \widetilde{\pi}\left(\boldsymbol{\eta}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}\right) \widetilde{\pi}\left(\boldsymbol{\theta}^{(k)}|\mathbf{y}\right) \Delta_k \,, \qquad (3)$$

where $\widetilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is as in Eq. (2) and $\widetilde{\pi}(\boldsymbol{\eta}_i|\boldsymbol{\theta})$ is an approximation of $\pi(\boldsymbol{\eta}_i|\boldsymbol{\theta})$. INLA provides three primary methods for computing $\widetilde{\pi}(\boldsymbol{\eta}_i|\boldsymbol{\theta})$, termed the *Gaussian*, *Laplace*, and *Simplified Laplace* strategies, in addition to *adaptive* and *automatic* strategies. Each strategy applies Laplace approximations or series expansions to different conditional distributions, and has different trade offs for efficiency and accuracy. For full details on these methods, see for example [44, 53, 54].

Predictions in INLA are carried out concurrently with model fitting, where the posterior predictive distribution of the response at each prediction location is computed [53]. The INLA software provides summary statistics including the mean, median, standard deviation and quantiles of the predictive distribution.

INLA-based model

A model using the INLA software is formulated similar to [55] and [56] to predict malaria prevalence. Let H_i denote the number of positive results (e.g., in this case, malaria infections) and N_i the number of people screened at location $\mathbf{x}_i$ for $i = 1, \ldots, n$. Let p_i denote the modelled prevalence at location $\mathbf{x}_i$, and $\mathbf{p}$ be the vector of modelled prevalences over all locations. Then H_i is modelled using a binomial distribution as

$$H_i \sim \text{Binomial}(N_i, p_i).$$

The standard link for the binomial distribution is the logit function, which opens-up the probabilities in $[0, 1]$ to real values in $(-\infty, \infty)$. Thus obtaining,

$$\text{logit}(p_i) = \beta_0 + S(\mathbf{x}_i) \,, \qquad (4)$$

where β_0 denotes the intercept and S is a spatial random effect that follows a zero-mean Gaussian process with Matérn covariance function:

$$\text{cov}\big(S(\mathbf{x}_i), S(\mathbf{x}_j)\big) = \frac{\boldsymbol{\sigma}^2}{2^{\boldsymbol{\lambda}-1}\Gamma(\boldsymbol{\lambda})} \big(\kappa||\mathbf{x}_i - \mathbf{x}_j||\big)^{\boldsymbol{\lambda}} K_{\boldsymbol{\lambda}}\big(\kappa||\mathbf{x}_i - \mathbf{x}_j||\big) \,. \qquad (5)$$

Here $\boldsymbol{\lambda}$ is the smoothness parameter, $\boldsymbol{\sigma}^2$ denotes the variance and $K_{\boldsymbol{\lambda}}$ is the modified Bessel function of the second kind. The parameter κ controls how fast the correlation decays with distance.

The implementation and parameter settings for the model are based on the examples available in [56]. The first step in setting up a model is to construct a triangular mesh on which the SPDE will be solved. The software constructs this mesh based on restrictions provided by the user, and it usually contains a region of smaller triangles near the data surrounded by an extension of coarser triangles to avoid boundary effects [57]. When using the Kenya data, the maximum triangle edge length was set to 0.5 for the inner region, and 4 for extension. The `cutoff` parameter sets a distance, under which, points are grouped together when constructing the mesh vertices. This has been set to 0.01, and additionally the `min.angle` and `offset` parameters, which determine the minimum allowed angles in the triangles and the size of the extension, have been left at their default values of 21 degrees and -0.1, respectively. When using the Africa data, a mesh on the unit sphere is used, with the above parameter values converted to radians.

The smoothness parameter $\boldsymbol{\lambda}$ in the Matérn covariance function Eq. (5) must be chosen via the `alpha` parameter

$$\boldsymbol{\lambda} = \alpha - \frac{d}{2} \,,$$

where d is the dimension of the space (i.e. 2 for a spatial model). The `alpha` parameter has been set to its default value of 2.

User settings additionally control the approximations during inference. The default `auto` strategy was used for approximating $\widetilde{\pi}(\boldsymbol{\eta}_l|\boldsymbol{\theta}, \mathbf{y})$. The `int.strategy` parameter then determines how the points $\boldsymbol{\theta}^{(k)}$ are selected for the numerical integration in Eq. (3), and the faster *empirical Bayes* strategy was chosen. This selects a single point, namely the mode of $\widetilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and therefore does not average predictions over uncertainty in the hyperparameters, as would typically happen in an MCMC inference procedure. The median of the predictive distribution was used for point predictions, though other quantities such as the mean are available.

GPBoost

GPBoost combines tree-boosting with Gaussian processes and mixed effects models. Inspired by the performance of gradient boosting algorithms, as implemented in popular software such as XGBoost and LightGBM, GPBoost aims to leverage the advantages of tree-boosting algorithms including accounting for complex nonlinearities, discontinuities and higher order interactions with the versatility of Gaussian processes [46]. It has the functionality to use mixed effects models, in particular models with grouped random effects.

The general structure for a model which can be implemented using GPBoost is

$$Y_i \sim N\left(z_i, \boldsymbol{\tau}^2\right) \quad \text{for} \quad i = 1, \ldots, n,$$

$$\mathbf{z} = F(X) + Z\mathbf{S},$$

$$\mathbf{S} \sim \mathcal{N}(0, \Sigma), \tag{6}$$

where Y_i is the response variable at location $\mathbf{x}_i$. The matrix $X \in R^{n \times p}$ is the fixed effect predictor matrix, with the ith row containing covariates for location $\mathbf{x}_i$. The fixed effects function of the covariates F, is nonlinear and is learned with boosting. $\mathbf{S} \in R^m$ contains the random effects with covariance matrix $\Sigma \in R^{m \times m}$, while $Z \in R^{n \times m}$ is the random effect predictor variable matrix, which is typically used to define grouped random effects.In a Gaussian process model the random effects $\mathbf{S} = (S(\mathbf{x}_1), S(\mathbf{x}_2), \ldots, S(\mathbf{x}_m))$ are a finite-dimensional version of a Gaussian process $S(\mathbf{x})$ with a covariance function:

$$\operatorname{cov}\big(S(\mathbf{x}), S\big(\mathbf{x}'\big)\big) = c\big(\mathbf{x}, \mathbf{x}'\big)\,,\ \mathbf{x}, \mathbf{x}' \in R^d\,.$$

Here c is a covariance function often parameterized as:

$$c\big(\mathbf{x}, \mathbf{x}'\big) = \boldsymbol{\sigma}_1^2 r\big(||\mathbf{x} - \mathbf{x}'||/\boldsymbol{\rho}\big)\,,$$

where r is an isotropic autocorrelation function with $\boldsymbol{\sigma}_1^2 = \operatorname{var}(S(\mathbf{x}))$ and $\boldsymbol{\rho}$ is the range parameter which determines how quickly r decays with distance. GPBoost currently supports the exponential, Gaussian, Matérn, powered exponential, Wendland, and tapered exponential covariance functions. In a Gaussian process model, Z is usually encoded as a diagonal matrix, so that each element of $\mathbf{S}$ contains the spatial random effect for that location.

With its default settings, GPBoost does not apply approximations to the Gaussian process. For increased efficiency, Vecchia approximations are available in the software. These approximations assume conditional independence between responses based on their locations, resulting in sparse Cholesky factorizations of the precision matrix and in turn improved computational efficiency [46, 58].

Inference with GPBoost is carried out by jointly optimizing the nonlinear fixed effects function F, and the variance and covariance parameters $\boldsymbol{\theta}$ (i.e $.\boldsymbol{\tau}^2, \boldsymbol{\sigma}_1^2$, and $\boldsymbol{\rho}$). In the Gaussian process case, the goal of the optimization is to minimize the *risk functional:*

$$R(F, \boldsymbol{\theta}) = L\big(\mathbf{y}, F(X), \boldsymbol{\theta}\big)\,,$$

where $\mathbf{y} = \big(y_1, \ldots, y_n\big)$ are the observed responses at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$.Here, $L\big(\mathbf{y}, F(X), \boldsymbol{\theta}\big)$ is the negative log marginal likelihood for obtaining the observed responses $\mathbf{y}$, given the observed covariate matrix X, and model parameters $\boldsymbol{\theta}$,

$$L\big(\mathbf{y}, F(X), \boldsymbol{\theta}\big) = \frac{1}{2}\big(\mathbf{y} - F(X)\big)^T \Psi^{-1}\big(\mathbf{y} - F(X)\big) + \frac{1}{2}\log\det(\Psi) + \frac{n}{2}\log(2\pi)\,,$$

where $\Psi = Z\Sigma Z^T + \boldsymbol{\tau}^2 I$. The risk functional is minimized by iteratively updating F and $\boldsymbol{\theta}$. At step k, F_{k-1} is held fixed and $\boldsymbol{\theta}_k = \operatorname{argmin}_{\boldsymbol{\theta}}\big(L\big(\mathbf{y}, F_{k-1}(X), \boldsymbol{\theta}\big)\big)$ is computed using a gradient or quasi-Newton method. With this value of $\boldsymbol{\theta}_k$, F is updated via a single step of a boosting algorithm.

After optimization, GPBoost produces predictions in a similar manner to Gaussian process regression. The joint distribution of the observed and predicted responses is formed, and conditioned on the observed responses. The mean of the resulting conditional distribution is used for the predicted value of the response.

As covariates are not used in this implementation, tree boosting is used only to find the intercept. While this does neglect GPBoost's functionality for learning nonlinear functions of covariates, GPBoost has been included in the analysis for users who may wish to apply it in more complicated scenarios that may benefit from tree boosting.

GPBoost-based model

The basic model in Eq. (6) can be extended to use non-Gaussian likelihoods, where the implementation in the software uses a Laplace approximation during inference [59]. Currently Bernoulli-probit, Bernoulli-logit, Poisson, and Gamma distributions are supported for the response variable. However unlike INLA and FRK, a binomial response is not currently supported which does present a limitation for applying this method for mapping malaria prevalence. This work therefore models malaria prevalence by customizing Eq. (6) as follows:

$$H_i/N_i \sim N\Big(z_i, \boldsymbol{\tau}^2\Big)\,,$$

$$z_i = \beta_0 + S(\mathbf{x}_i)\,,$$

where β_0 is the intercept, and H_i and N_i denote the number of positive results and the number of people tested at location $\mathbf{x}_i$. Note that for simplicity, the direct proportion of positive tests was used rather than the empirical logit, and predictions were clipped to lie within $[0, 1]$ for the prevalence maps. The exponential covariance function $r\big(||\mathbf{x} - \mathbf{x}'||/\boldsymbol{\rho}\big) = \exp\big(-||\mathbf{x} - \mathbf{x}'||/\boldsymbol{\rho}\big)$ was selected, which is the default choice in the software. Notably, this model does not use GPBoost's full capability for learning nonlinear functions of the covariates, however it has been constructed in order to be consistent

with the choice to not use covariates for any of the models.

The parameter settings for the implemented model follow examples by the package author [60]. For the spatial random effect, a full Gaussian process without approximation was used, with the `gp_approx` parameter set to its default value of `none`. Other parameters in the software control the trees and boosting algorithm used to learn the fixed effects function F. The number of boosting rounds was set to 247 and the learning rate to 0.01, using the parameters `nrounds` and `learning_rate`. Other settings for the model include `num_leaves`=1024, `max_depth`=6, and `min_data_in_leaf`=5, each of which control the size of the trees.

SpRF

Spatial Random Forests (SpRF) [47] extend classical random forests to a spatial domain by using distances to observation points as explanatory variables, i.e. when fitting a model with SpRF, for each point $\mathbf{x}_i$, where y_i is given, covariates are used that give the distance from each other observation point. That is, the design matrix for this part of the model is simply the distance matrix between all pairs of observation locations. In order to obtain uncertainty estimates, the SpRF authors use quantile regression forests which estimate specified quantiles of the conditional distribution $Y_i|X_i$ [61] where X_i are the covariates for the ith response, in contrast to classical random forests which do not provide uncertainties.

The generic equation of an SpRF-based model is given by

$$Y_i = f\left(X_{G_i}, X_{R_i}, X_{P_i}\right),$$

where Y_i is the response at location $\mathbf{x}_i$, X_{G_i} denotes a vector of the distances to each of the observation locations from the querying point $\mathbf{x}_i$ (including a distance of 0 to itself, in the ith position of the vector) and X_{R_i} and X_{P_i} denote two types of covariates—surface reflectance and process-based. The function f is learned by the random forest. Unlike the other methods examined, SpRF does not use a covariance function.

SpRF is based on the `ranger` package for random forests, which provides an implementation of quantile regression forests with training procedure outlined in [61]. Point predictions are given by the estimated medians from the quantile regression forests.

SpRF-based model

As in [47], an additional normal assumption for the response was included in order to construct the simple SpRF model

$$H_i/N_i \sim N\left(z_i, \boldsymbol{\tau}^2\right),$$

$$z_i = f\left(X_{G_i}\right) \quad \text{for} \quad i = 1, \ldots, n,$$

where H_i and N_i are as defined above and X_{G_i} contains the distances from each observation point to $\mathbf{x}_i$.

The user parameters for SpRF determine the structure of the random forest and the rules for growing each tree, including the number of trees and the number of variables to split on at each node via the `num.trees` and `mtry` parameters. Each parameter was left at its default value, resulting in a forest with 500 trees where each node splits at $\sqrt{n_v}$ variables (n_v is the total number of variables input into the random forest). Other parameters which further tune the structure of the trees and forest have been left at their default values, and the code for the SpRF model is based on a tutorial from the method's authors [62].

FRK

Fixed Rank Kriging (FRK) [48] is a spatio-temporal modelling framework built for large datasets. It uses a spatial random effects (SRE) model, which decomposes a spatially correlated mean-zero random process using a linear combination of spatial basis functions. This dimensionality reduction using a relatively small number of basis functions ensures FRK's computational efficiency. The spatial domain D is partitioned into M subsets, $A_1, \ldots, A_M$, called basic areal units (BAUs) with centroids $\mathbf{x}_1, \ldots, \mathbf{x}_M$. The SRE model is constructed on these BAUs which determine the granularity of the model, and the process is assumed to be piecewise constant over the BAUs.

The general equation for a model implemented in FRK with a Gaussian response can be written as

$$Y_i \sim N\left(z_i, \boldsymbol{\tau}^2\right) \quad \text{for} \quad i = 1, \ldots, n,$$

$$\mathbf{z} = C_Z \zeta,$$

$$\zeta_j = t\left(\mathbf{x}_j\right)^T \beta + \nu\left(\mathbf{x}_j\right) + \xi\left(\mathbf{x}_j\right) \quad \text{for} \quad j = 1, \ldots, M.$$

Here, $Y_i, i = 1, \ldots, n$ are the responses at the observation locations, $\zeta = (\zeta_1, \ldots, \zeta_M)^T$ is the value of a latent spatial process evaluated at each of the BAUs with centroids $\mathbf{x}_1, \ldots, \mathbf{x}_M$, and C_Z is an n by M matrix connecting the observation locations to the BAU locations. The vector $\mathbf{t}\left(\mathbf{x}_j\right)$ is a collection of covariates at BAU j and $\boldsymbol{\beta}$ is a vector of regression coefficients, while $\nu\left(\mathbf{x}_j\right)$ is the value of a small-scale, spatially correlated random effect. Lastly, $\xi\left(\mathbf{x}_j\right)$ is a fine-scale random effect, which is treated as uncorrelated across the BAUs [48].

FRK introduces non-Gaussian data to the model by replacing the observation distribution with a member of the exponential family and using a link function to transform the latent process into a mean process [63]. The general structure of such a model is

$$Y_i|\boldsymbol{\mu}_i, \psi \sim \text{EF}(\boldsymbol{\mu}_i, \psi) \quad \text{for } i = 1, \ldots, n,$$

$$\boldsymbol{\mu} = C_Z \boldsymbol{\mu}',$$

$$g(\boldsymbol{\mu}') = \zeta,$$

$$\zeta_j = \mathbf{t}(\mathbf{x}_j)^T \boldsymbol{\beta} + \nu(\mathbf{x}_j) + \xi(\mathbf{x}_j) \quad \text{for } j = 1, \ldots, M, \quad (7)$$

where ψ is a dispersion parameter for the context dependent member of the exponential family EF, $\boldsymbol{\mu}$ is called the *mean process*, and $g(\cdot)$ is the link function. The mean process at the observation locations is represented by $\boldsymbol{\mu}$, while $\boldsymbol{\mu}'$ represents the mean process at the BAUs.

The spatially correlated random effect $\nu(\mathbf{x})$ is decomposed as

$$\nu(\mathbf{x}) = \sum_{l=1}^{r} \phi_l(\mathbf{x}) \boldsymbol{\eta}_l,$$

where $\phi_1, \ldots, \phi_r$ are a fixed collection of basis functions on the spatial domain, and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_r)^T$ is an r-variate Gaussian random variable with covariance matrix K. To estimate model parameters including the coefficients $\boldsymbol{\beta}$, variance parameters for the fine scale random effect ξ, and covariance parameters for the covariance matrix K, FRK carries out maximum likelihood estimation. When working with non-Gaussian data, a Laplace approximation is used to approximate the marginal likelihood, which is then maximized via a quasi-Newton method.

By default, FRK produces a prediction for the mean process $\boldsymbol{\mu}(\cdot)$ at each of the BAUs. Predictions and uncertainties are generated via a Monte Carlo sampling approach, and the predicted value of $\boldsymbol{\mu}$ in each BAU is taken to be the average of the samples.

FRK-based model

As with INLA, the number of positive tests H_i is modelled using a binomial distribution

$$H_i \sim \text{Binomial}(N_i, p_i),$$

$$\mathbf{p} = C_Z \mathbf{p}',$$

where p_i is the prevalence at the ith observation location. The vector $\mathbf{p}'$ gives the prevalence at the BAUs, and is transformed into the prevalence at the observation locations via the C_Z matrix, which has construction detailed in [63]. The logit function is then used as the link function g in Eq. (7), i.e.

$$\text{logit}(p_j') = \zeta_j.$$

As covariates are not being used, the latent process over the BAUs ζ_j can be written as:

$$\zeta_j = \beta_0 + \nu(\mathbf{x}_j) + \xi(\mathbf{x}_j),$$

where β_0 denotes the intercept.

The implemented model decomposes the spatial random effect, $\nu(\mathbf{x})$, using Gaussian basis functions of two different scales placed regularly across the spatial domain, as controlled by the `type`, `nres`, and `regular` parameters respectively. The spatial scale of these basis functions is determined jointly by the `regular` parameter and the `scale_aperture`, which have been left at their default values of 1 and 1.25 respectively. The assumed correlation structure of the random coefficients $\boldsymbol{\eta}$ is controlled by the `K_type` parameter. When using a non-Gaussian model, this takes a default value of `precision`, which models the coefficient dependence using a precision matrix Q based on the Leroux model [63]. During prediction, the user can specify the number of Monte Carlo samples to be drawn, which has been left at the default value of 400. Code and parameter choices for the implemented FRK model are based on examples from the package authors in [48, 63].

Methods for the country scale analysis

At the country scale, the models implemented using the four methods were compared qualitatively using their predictive maps, while cross-validation was used to compare their predictive performance. To produce maps of predicted prevalence, each model was fit on all available *P. falciparum* prevalence surveys from Kenya in 2009 from the malariaAtlas R package. This consisted of 382 surveys at points across the country which are shown in Fig. 1a. Point estimates of prevalence and uncertainties were produced by the fitted models on a grid over Kenya, with each cell covering a nominal 0.1 degrees (approximately 11×11 km at the equator) in longitude and latitude.

Model fitting and prediction were run on a 2014 MacBook Pro with a two core, 2.8 GHz Intel Core i5 processor running macOS 10.13.6. Each model was run using a single thread to obtain a baseline performance comparison to accompany the model predictions, although parallelization options are available for each model which may provide performance improvements.

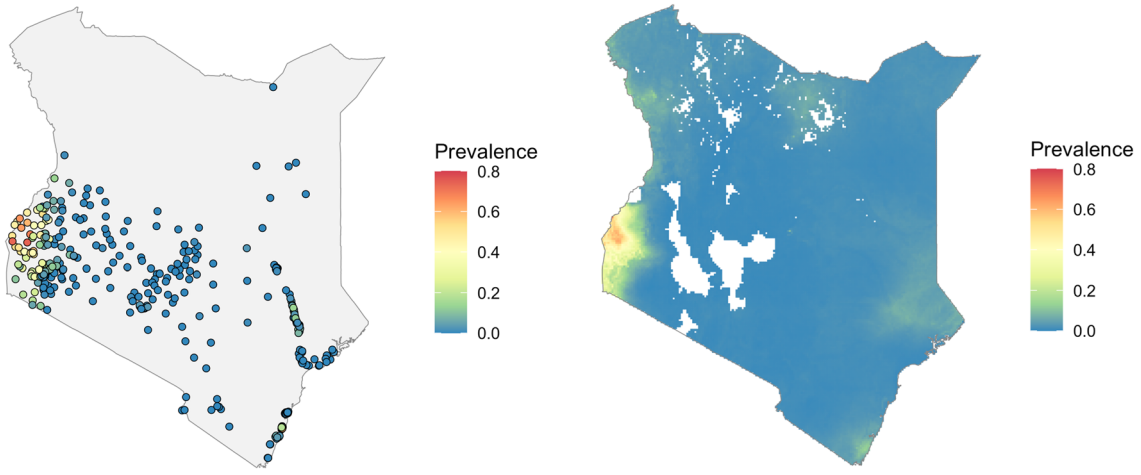

Fig. 1 2009 *P. falciparum* prevalence data in Kenya. **a** shows prevalence survey results, while **b** shows the Malaria Atlas Project predicted prevalence

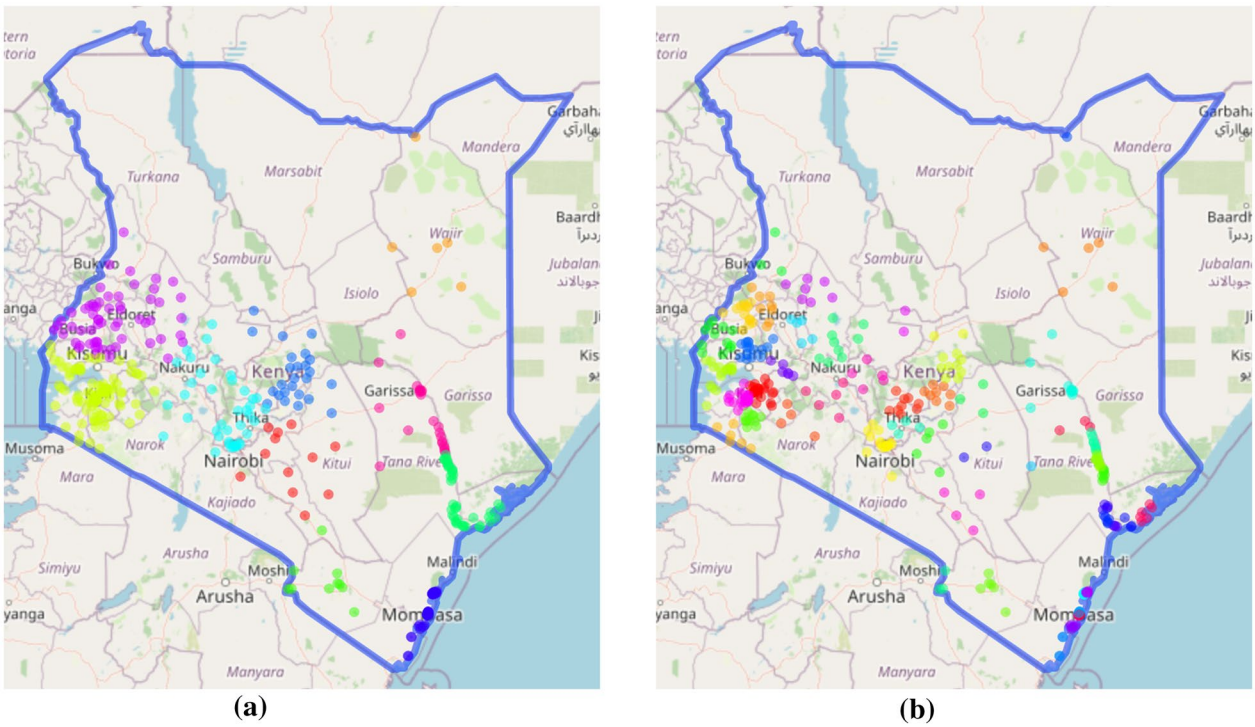

Fig. 2 *P. falciparum* prevalence survey locations in Kenya for 2009. Colours represent different cross-validation folds. **a** and **b** show 10-fold and 50-fold cross-validation locations respectively

Recorded times were measured as the total time to run a model's R script, including both fitting and prediction.

To evaluate the models, spatial block cross-validation (CV) [64] was used with both 10 and 50 folds. In a spatial setting, randomly allocating points to cross-validation folds is not effective because close by points can act as proxies. The folds were selected using *k*-means clustering [65] on the spatial coordinates of the prevalence surveys—resulting in a series of 'blocks' of spatially-adjacent points. Figure 2 shows the location of points for the two

sets of CV folds, where each colour represents a fold. The 10 and 50 CV folds measure different abilities of the methods. The 50-fold CV quantifies short-scale interpolation ability, while the 10-fold quantifies the ability to interpolate over longer distances.

Using 10 and 50-fold cross-validation, the following are investigated:

1. analysis of the point predictions including a comparison between the predictions and out-of-sample prevalence values using multiple measures,
2. analysis of the uncertainty bounds for each model, and
3. analysis of the predictions with respect to density of the sampled locations ("Detailed cross-validation results").

Analysis of uncertainties is complicated by the differing measures of uncertainty provided by the different methods. INLA contains information on the summaries of the posterior marginal densities of the fitted model, and can compute the standard deviation and different quantiles of the predictions. GPBoost provides the variance of each prediction. FRK predicts the standard deviation of each prediction in the linear, Gaussian setting. For the non-Gaussian case, it provides the predictions using a Monte Carlo approach [63]. SpRF uses quantile regression and the quantiles can be specified in the ranger package. To compare SpRF with the other methods, a normally distributed response is assumed as in [47], and the standard deviation for SpRF's predictions is estimated as

$$\mathrm{SD} \approx \mathrm{IQR}/1.34898.$$

Hengl et al. note that this assumption may not always be valid [47], and hence only a rough comparison of the SpRF model's uncertainty with the other three models is possible.

For each model, the number of observed prevalence values which lie within the predicted uncertainty intervals is measured. Let $\widehat{y_i}$ denote the mean of the predicted response for observation y_i, and define

$$\text{Within 1SD}\left(\widehat{y_i}\right) = \text{TRUE if } \left|y_i - \widehat{y_i}\right| \leq \text{SD}\left(\widehat{y_i}\right)$$

$$\begin{aligned}\text{Within 2SD}\left(\widehat{y_i}\right) &= \text{TRUE if Within 1SD}\left(\widehat{y_i}\right) \\ &= \text{FALSE and } \left|y_i - \widehat{y_i}\right| \leq 2\text{SD}\left(\widehat{y_i}\right)\end{aligned}$$

where SD denotes the standard deviation. As prevalence values are between 0 and 1, the bounds are trimmed if they exceed these limits. Note that $\widehat{y_i}$ corresponds to the predicted prevalence for the implemented GPBoost and FRK models, but not for the INLA and SpRF models which use the median for predictions.

Methods for the continent scale analysis

At the continent scale, analysis focused on the prediction maps and each model was fit three sets of prevalence data over Africa. The first set consists of 868 *P. falciparum* prevalence surveys from 2009, available via the `malariaAtlas` R package. This data is shown in Fig. 3a, with survey points concentrated in Kenya and Somalia. Each model was additionally fit using two types of simulated data to allow comparison of the predictions with a known truth and to compare model performances on both interpolation and extrapolation tasks, and lastly to assess how properties of the data such as spatial sparsity and noise impact model predictions.

Simulated data was generated using the 2009 *P. falciparum* prevalence raster created by the Malaria Atlas Project (MAP), shown in Fig. 3d [13]. Prevalence was sampled from the raster at the locations of the 2009 surveys and combined with the number of tests at each location to generate a binomial sample for the number of people testing positive. Of the 868 observation locations, 28 points lie on gaps in the prevalence raster and were excluded, leaving 840 points in this second dataset, which is shown in Fig. 3b. This spatially clustered simulated data allows for the evaluation of each model's ability to extrapolate over regions with little or no data. While this dataset shares locations with the observation data, its prevalence notably contains less noise.

The second set of simulated data was generated by selecting 1000 points at random on the MAP raster, allowing for comparison of the models' interpolation performance when trained on data with good spatial coverage. A binomial sample for the number of positive tests was generated at each location, where the number of people tested was set to 85, approximately the average number in the surveys from 2009. The prevalences from this simulated dataset are shown in Fig. 3c.

The same parameter settings were used as in the country scale analysis, though whenever possible, settings that compute an appropriate spherical distance between points were chosen due to the larger spatial extent of the data. This was possible SpRF which uses great circle distances, and for INLA which allows for meshes to be constructed on the unit sphere. GPBoost does not have this functionality at the time of writing this article, however correspondence with the package authors reveals that they hope to add this functionality in future. While FRK does support using great circle distances for some models, this feature is not currently well supported for models with a binomial response and did not work in implemented tests. Hence, this work

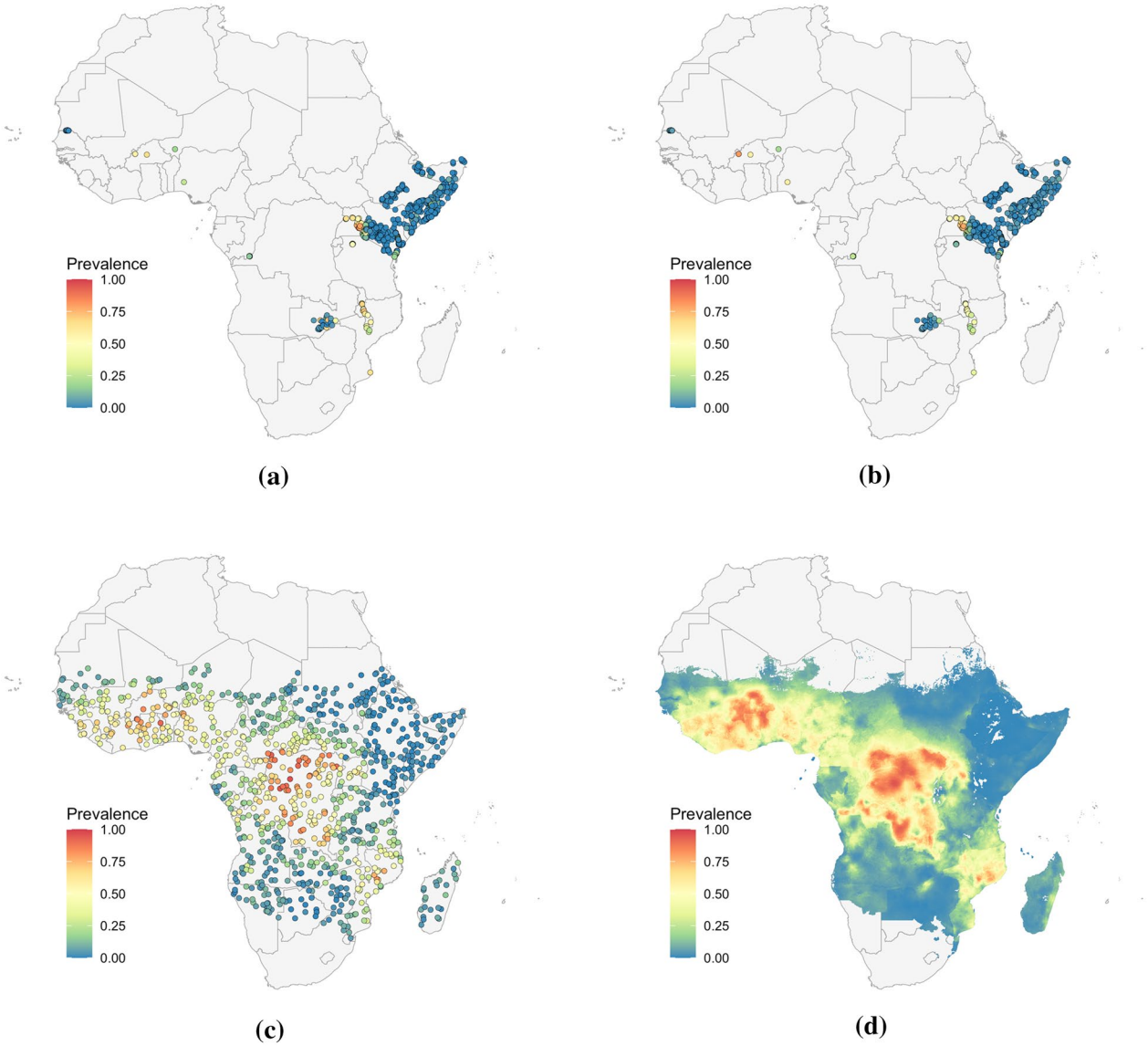

Fig. 3 *P. falciparum* prevalence data used to fit the four models at the continental scale. **a** shows the 2009 observed data at 868 locations. **b** shows the prevalence generated from binomial samples at the observation locations. **c** shows the prevalence generated by binomial samples at 1000 uniformly random locations. **d** is the Malaria Atlas Project predicted prevalence raster from 2009 used to generate the samples in **b** and **c**

uses Euclidean distances between coordinates for both GPBoost and FRK.

Using INLA with a spherical geometry requires a mesh to be built on a subset of the sphere. Although several methods for constructing this mesh are used in the literature [52, 57, 66], each produced similar results and hence the method outlined by Lindgren and Rue [57] was selected.

Model fitting and prediction were carried out on a single 3.00 GHz Intel(R) Xeon(R) Gold 6154 CPU in the Physical partition of the University of Melbourne's high performance computing cluster, Spartan, and each model run was allocated 32 GB of RAM. As with the country scale data, each model was run using a single thread. Predictions were produced on a grid with cell side length 0.15 degrees (approximately 16.7 km at the equator).

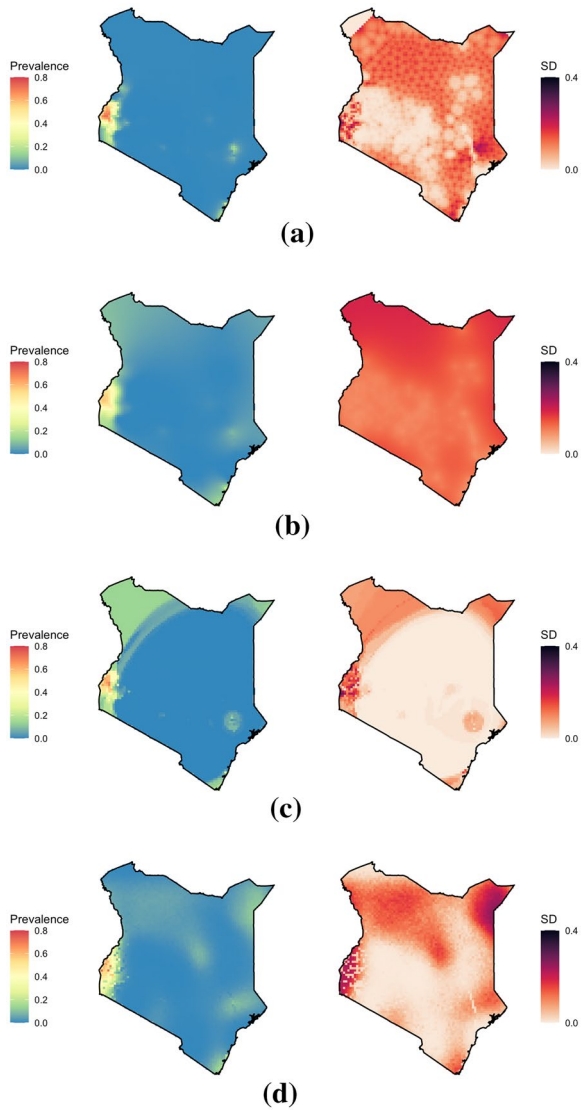

Fig. 4 Predicted prevalences and uncertainties for **a** INLA, **b** GPBoost, **c** SpRF, and **d** FRK when trained on *P. falciparum* prevalence data from Kenya in 2009. Note that these maps are intended only to illustrate differences in model predictions when fit to a small data sample, and are not likely to accurately represent malaria prevalence across the country in this year

Results

This section presents the analysis on Kenya, which includes the predictive maps and cross-validation results, and the continent scale analysis including models trained on three different input datasets discussed above.

Case study: Kenya

At a national scale, two means of verifying the implemented models have been used: 1. predictive maps, and 2. 10-fold and 50-fold cross-validated predictions. The whole dataset was used to produce the predictive maps, while for cross-validation, some data was left out in each fold.

Predictive maps

The predictions and uncertainties produced by the four models when trained on the 2009 Kenya prevalence data are shown in Fig. 4 At the broadest scale, each model is similar in predicting a region of high prevalence in Western Kenya, with clusters of higher prevalence in the East, but low prevalence over much of the rest of the country. For each model, the predicted prevalence drops to zero quite quickly away from the data, indicative of a smaller spatial range than might be expected. This is especially prominent with the INLA-based model, and may be indicative of overdispersion in the data.

A notable feature is the arc like band of higher prevalence in the north west of Kenya in SpRF's predictions in Fig. 3c, which is further discussed in "Continent scale results" section. Higher prevalence in this region is also somewhat apparent in the GPBoost-based model's predictions and, to a lesser extent, FRK. This area of predicted higher prevalence falls in a broad region with no prevalence data and so represents different approaches to extrapolation in the four models.

Cross-validation results

Table 1 gives the cross-validation results. In terms of cross-validation RMSE and correlation, FRK performs the best for 10-fold CV, and GPBoost performs the best for 50-fold CV. SpRF predictions had the highest correlations to the data used to train the model, but poorer correlation to out-of-sample data, indicating that

Table 1 Cross-validation results of the four models

Model	10-fold- RMSE	50-fold RMSE	Training Correlation	10-fold Correlation	50-fold Correlation	% points within 1SD (10-fold)
INLA	0.181	0.124	0.909	0.235	0.683	75
GPBoost	0.127	**0.11**	0.873	0.646	**0.751**	84.211
SpRF	0.132	0.121	**0.912**	0.641	0.702	37.105
FRK	**0.125**	0.123	0.902	**0.661**	0.702	83.421

Boldface denotes the best score in each column

this model may be overfitting to the training data. INLA performs poorly with respect to the 10-fold RMSE and correlation.

Table 1 shows that SpRF has only 37.105% of the points within 1SD for 10-fold cross-validation, which is much lower than for the other models and is discussed further in "Detailed cross-validation results" section. GPBoost performs the best in terms of the percentage of points within 1SD. However these results need to be taken in context, because a higher standard deviation can increase this percentage.

Figure 5 shows the interval and point predictions for the four methods for 10-fold cross-validation. Points within 1SD are shown in green, points within 2SD are shown in blue, and the rest are shown in red. Many of INLA's predictions are close to zero and this issue is investigated in "Effects of input noise on INLA" section. Additional cross-validation metrics that consider the prediction error divided by its standard deviation are detailed in [67]. However, this work does not consider these metrics as some methods produce very small standard deviations and thus will result in very large values. More details on cross-validation results in terms of the clusters and density of locations are given in "Detailed cross-validation results" section.

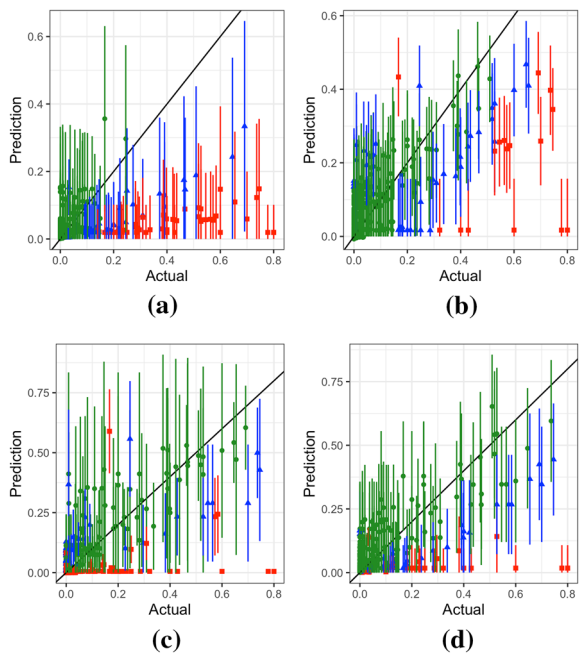

Fig. 5 Interval predictions for 10-fold cross-validation for **a** INLA, **b** GPBoost, **c** SpRF, and **d** FRK using the national level Kenya data. Points show the predicted mean from each model, and intervals show one standard deviation above and below the mean

Continent scale results

Geostatistical mapping is often carried out at a continent or global scale and frequently uses large datasets of observations. As computation time for inference and prediction can scale poorly with the amount of data and size of the domain, it is important to assess the performance, both in terms of predictive power and time, of recent methods. To examine how each of the four methods perform at larger scales, and to understand how predictions are affected by potential violations of model assumptions and by clustering and sparsity of observation points, the study area was expanded to the whole continent of Africa in 2009, and each model was fit to the three prevalence datasets shown in Fig. 3. Figure 6 shows the prevalence predicted by the models when trained on each input dataset. The corresponding uncertainties appear in "Prediction uncertainty".

Scaling up to the continent level reveals differences between the methods that are not apparent at smaller scales. While the national scale prevalence maps in Fig. 4 are largely similar apart from the slight banding effect seen in SpRF's predictions, the prevalence maps in Fig. 6 differ significantly, with artifacts appearing in several of the maps.

Overall, the four models are better at local interpolation than extrapolation over large regions without data. The predictions in Fig. 6(iii) generated using the randomly distributed data recover the prevalence structure of the MAP raster in Fig. 3d much more faithfully than the predictions in Fig. 6(ii) from the sparser non-uniform data. This behaviour is expected as malaria prevalence is known to be highly heterogeneous and the implemented models do not use covariate data.

SpRF's predictions display a prominent banding effect, visible in both the country and continent scale maps where contiguous arc-like bands of high prevalence appear in both point and uncertainty estimates. This may be explained by the fact that SpRF models the quantity of interest—malaria prevalence in this case—based on distances to points with known values. Thus bands of high or low prevalence appear at different radii from clusters of observations, and the piecewise constant nature of random forests would contribute to the sharp steps between each band. The banding effect is particularly prominent in Fig. 6(ci) and (cii), where the points were clustered into smaller regions, while it is less obvious when the datapoints have good spatial coverage, as in Fig. 6(ciii), which does not show bands spanning the continent. Further increasing the number of simulated points was found to further reduce the prominence of these bands.

Even though SpRF produces maps with this unwelcome feature, the cross-validated point estimates are quite

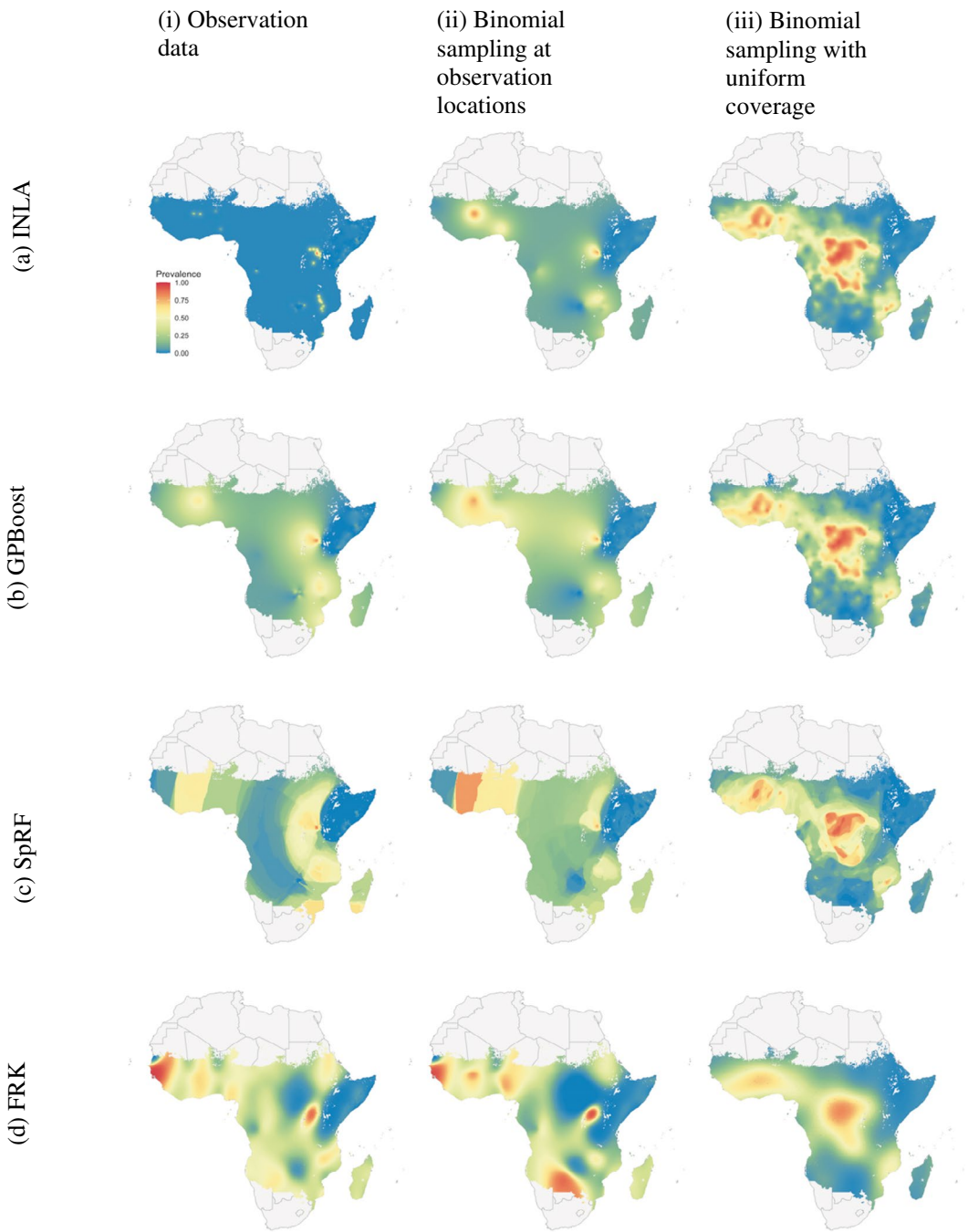

Fig. 6 *P. falciparum* prevalence predictions when fit using three different datasets. In column (i), models are fit using the survey data from Africa in 2009, shown in Fig. 3a. In column (ii), the models are fit to binomial samples drawn from the Malaria Atlas prevalence raster at the same survey locations, shown in Fig. 3b. In column (iii), they are fit to binomial samples drawn from the raster at 1000 uniformly selected locations across the continent, shown in Fig. 3c. Outputs have been masked by the Malaria Atlas Project raster in Fig. 3d. Note that these maps are intended only to illustrate differences in model predictions and are not likely to accurately represent malaria prevalence in this year

accurate. Table 3 in "Detailed cross-validation results" shows that SpRF has the highest proportion of points with absolute errors less than 0.05 and 0.1 for 10-fold cross-validation, which is a harder task for the algorithms than 50-fold cross-validation. Thus in this example, SpRF gives reliable predictions at points even though it may produce a predictive map that can be misleading in regions where there are no sample points.

Figure 6(ai), produced by INLA with the observation data, displays a sudden drop in prevalence away from

observations, resulting in flat near-zero predictions covering most of the continent. This appears to result from a combination of both the sparsity and noise present in the data, rather than clustered nature of the data alone. Figure 6(aii) uses nearly the exact same locations, yet shows higher values of prevalence spreading much further from the observations. "Effects of input noise on INLA" outlines evidence that this effect arises from unaccounted-for overdispersion in the observation data. In particular, increased noise in the data appears to reduce the estimated range for the spatial random effect, resulting in the model reverting to constant predictions away from observation locations. This behaviour is consistent with the INLA-based model's poor performance in the 10-fold cross-validation analysis in "Cross-validation results" section, where the model predicted near-zero malaria prevalence for each of the held out folds.

The FRK-based model's predictions depend strongly on the arrangement of the basis functions, which are generally placed by the software based on the data locations and the user parameters introduced in "FRK-based model" section. For example, FRK's predictions in Fig. 6(di) and (dii) display spurious oscillations in regions with little or no data, however these oscillations correspond to periodic placement of the basis functions. In regions with little data, FRK reverts to the prior mean, which varies with the basis functions and leads to oscillating predictions (A. Zammit Mangion, personal communication, September 1 2023). The spurious oscillations should generally coincide with locations where the predicted uncertainty is high, however notably Fig. 6 (di) and (dii) include a large patch of low prevalence over South Sudan where no data is located, yet where the predicted standard deviation is low in Fig. 20. Other tested arrangements led to flat predictions over the whole continent (results not shown). These types of artifacts are not present in Fig. 6(diii), where the input data has good spatial coverage. However, this map appears as a smoothed version of the input data, and does not resolve the finer structure in the MAP surface. The impact of the arrangement and number of basis functions on the prediction maps is detailed further in "FRK sensitivity" section.

For both Kenya and Africa, the GPBoost-based model produces prevalence maps without the artifacts appearing in the other models' outputs. However, the uncertainty maps in Fig. 20(ci)-(cii) exhibit a high level of overall uncertainty regardless of whether the regions have more survey points or not. This is further confirmed by the near-constant interval widths that rarely fluctuate with the density of the survey points in Fig. 16 ("Detailed cross-validation results" section). Even though GPBoost currently computes only Euclidean distances between coordinates, both the prevalence maps for Africa and Kenya appear to be reasonable. However, it is sub-optimal to use Euclidean distances between longitude and latitude coordinates for a global model.

Computational results

Times taken to train each model on each of the datasets and produce predictions are shown in Table 2. While FRK is consistently the fastest, INLA shows great variation among the African datasets, ranging from less than 10 min with the uniform simulated data to 69.11 min with the observation data. Further analysis of this variation for INLA is given in "Effects of input noise on INLA".

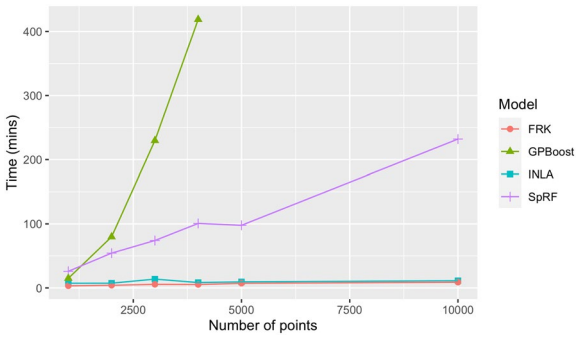

Fig. 7 Times taken by each model on uniformly distributed simulated datasets. GPBoost was not run with 5000 or 10,000 points due to the likely long computation time

Table 2 Times taken in minutes to train the models on each dataset and generate the prediction maps

		Dataset			
		Kenya: Observation	**Africa: Observation**	**Africa: Simulated observation**	**Africa: simulated uniform**
Model	INLA	0.34	69.11	11.49	7.05
	GPBoost	0.99	11.27813	6.56	13.85
	SpRF	0.44	24.54	24.64	27.6
	FRK	0.35	3.28	3.41	3.09

The Kenya: Observations column corresponds to the maps in Fig. 4. The Africa: Observations, Africa: Simulated observations, and Africa: Simulated uniform columns correspond to columns (i), (ii), and (iii) of Fig. 6 respectively. Note that different machines were used to run the models for the Kenya and Africa datasets

Each model was additionally trained on simulated prevalence datasets with 1000–10,000 points selected at random. Figure 7 shows times taken to fit each model and produce predictions as the dataset size varies. Both INLA and FRK remained very fast on larger datasets, showing little variation in their times. In contrast SpRF's time appears to increase linearly with dataset size, and GPBoost rapidly slows down on larger datasets, reflecting the computational requirements of using an unapproximated Gaussian process. As noted in GPBoost, a Vecchia approximation is available for this method to improve the computational efficiency. In "GPBoost with the Vecchia approximation", the effects on computation and model predictions of applying this approximation are examined.

Sensitivity of FRK and INLA to parameter choices

Of the four methods, INLA and FRK show promise in their computational efficiency, displaying favourable scaling compared to SpRF and GPBoost in Fig. 7. Additionally, while the artifacts in SpRF's output appear to stem from the way it uses distances as an input, it is less clear whether the artifacts in INLA and FRK's prevalence maps are due to specific model parameters, or if they are fundamentally caused by the approximations used by each method. For this reason, these two methods are examined more closely and test the sensitivity of their predictions to the model parameters.

INLA sensitivity

The primary artifact visible in the INLA-based model's prediction maps is the flat, near zero, predictions when the model is fit to the observation data, as shown in Fig. 6(ai). "Effects of input noise on INLA" outlines evidence that this feature is due to overdispersion, suggesting that the binomial response is unsuited for modelling the variability in the observed malaria data, despite commonly being used in tutorials on the application of INLA to disease mapping problems.

Several adjustments can be made to the model to address this overdispersion, such as the use of a Beta-binomial or Gaussian response (either directly on the proportion positive, or its empirical logit transform), or the inclusion of an independent error term in the linear predictor. All of these options include an additional parameter in the model to capture error variance at the level of the observation. Figure 8 shows predictions from an INLA-based model with a Beta-binomial response which has been fit to the observation data. The flat predictions of Fig. 6(ai) are notably absent, suggesting that the Beta-binomial is effective in resolving the overdispersion. A Gaussian response was additionally tested and was found to also handle the variability in the observation data, with results shown in "INLA with a Gaussian response". These results highlight a need for caution when applying INLA with a binomial response to disease mapping problems, and the importance of checking for overdispersion.

FRK sensitivity

While the fastest of the four methods, FRK's continent scale predictions display a spurious 'spotty' pattern when fit to either of the spatially sparse datasets and a much less detailed map when fit to the simulated data at randomly selected locations (Fig. 6). These features appear to stem from FRK's use of a small number of basis functions in approximating the Gaussian process. This section examines whether increasing the number of these functions can resolve the artifacts in FRK's outputs.

The number of basis functions used in FRK's approximation is primarily controlled by the `nres` and `regular` parameters. Increasing the `nres` parameter adds an additional 'resolution' or layer of basis functions with a finer spatial scale, while increasing the value of `regular` reduces the scale of each basis function and adds

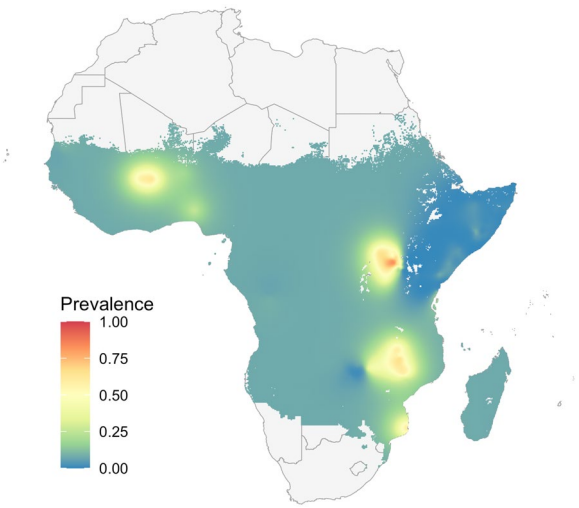

Fig. 8 Prevalence predictions from an INLA-based model with a Beta-binomial response, fit to the observation data in Fig. 3a

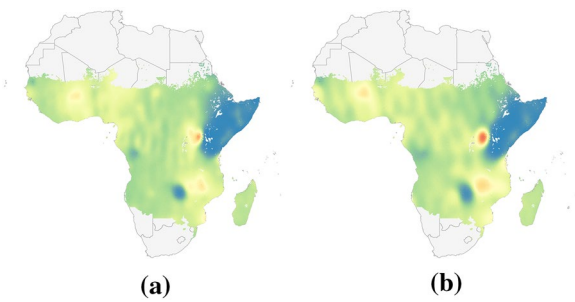

Fig. 9 *P. falciparum* prevalence predictions from the FRK model with **a** `nres` = 3 and `regular` = 1, and **b** `nres` = 2 and `regular` = 2

additional rows and columns to their arrangement. Details on the effects of these parameters are available in the software documentation [68]. The model used throughout "Case study: Kenya" to "Computational results" section had these parameters set to `nres` = 2 and `regular` = 1.

Figure 9a shows FRK's predictions when fit to the observation data with `nres` increased to 3, and `regular` left at 1, which resulted in a model with 1338 basis functions. Whilst the broad-scale 'spottiness' is less prominent in this model, finer-scale oscillation is quite visible in regions of Central Africa. This modest improvement came at a significant computational cost, as the model took over 55 min and required 106 GB of RAM, compared to the 4.77 GB of RAM and 3.28 min required when `nres` was set to 2, and `regular` was set to 1. Figure 9b shows the predictions when `nres` is kept at 2 and `regular` is increased to 2. These settings resulted in 600 basis functions, and required 22.74 GB of memory and 9 min to run, significantly less than when increasing the `nres` parameter. However the fine-scale oscillation is noticeably more pronounced in areas with little or no data.

While the key to FRK's computational efficiency is its decomposition of the spatial random effect into a small number of basis functions, these results suggest that it is challenging in practice to balance this efficiency with the risk of artifacts appearing in the model output, especially on large scale mapping problems, with sparse data. The observed issues with computational efficiency may result from the current software implementation rather than from FRK's underlying mathematical approach (A. Zammit Mangion, personal communication, September 1 2023), and hence such a balance may be easier to achieve in future software versions.

Discussion

This applied comparison of four computational spatial modelling methods found that two of them (SpRF and GPBoost) are not sufficiently scalable or accurate to be applicable to large-scale malaria prevalence modelling problems. SpRF's spatial predictions displayed a prominent 'banding' artifact, and at first glance the SpRF-based models appeared to be overfitted (matching closely to training data, but making poor predictions to hold-out data). However, closer inspection ("Detailed cross-validation results") shows that SpRF's tighter uncertainty intervals in low density regions may result in this perception. Unlike the other methods, SpRF does not incorporate a covariance function, but instead treats the columns of the distance matrix between coordinates as covariates for inclusion in the Random Forest. A covariance function could in fact be applied to the distance matrix before inclusion in the model. This would not have the same interpretation as in the Gaussian-process based models considered in this work, but would enable SpRF to consider a distance-based decay in the unobserved spatial effects being modelled. However, the Random Forest inference machinery in SpRF would have no means to estimate the parameters of such a function (such as the rate of decay with distance), and it seems unlikely they could be reasonably specified in advance. Due to these issues of fit, and the fact that the computation time of SpRF scaled approximately linearly with the size of the data, this approach is unlikely to be useful for applied spatial modelling of malaria data.

The GPBoost-based model made reasonably good predictions to hold-out data, being the best-performing model at 50-fold spatially-blocked cross-validation in the national-scale comparison (implying a good ability to extrapolate over short distances) and the second-best, behind FRK, at 10-fold cross-validation (ability to extrapolate over longer distances). However the computation time using the default GPBoost specification scaled very poorly with increasing data size. This is because by default GPBoost performs inference on the full (unapproximated) Gaussian process, with each step of the inference procedure requiring an $\mathcal{O}(n^3)$ inversion of the covariance matrix. Neither the maximum-likelihood inference of GP hyperparameters, and boosting inference on the intercept (and covariate effects if used) reduce this computational burden. Employing the Vecchia approximation available for the method did not resolve all issues, as shown in "GPBoost with the Vecchia approximation". The Vecchia approximation resulted in faster computation times, but also resulted in artifacts in the model predictions. Increasing the complexity (number of neighbouring points to consider) in the approximation reduced these artifacts, but at the cost of a substantial increase in the required computation time and RAM usage. It is worth noting that GPBoost is a relatively new technique, and future versions may include faster approximations.

Both INLA and FRK offered substantially better scalability to increasing data size than SpRF and GPBoost, taking only minutes to fit to 10,000 datapoints. Whilst it was computationally scalable, and is a widely established method and software for geostatistical modelling of malaria data, implementing INLA using the commonly suggested binomial distribution for prevalence data (e.g. as suggested in [31, 56]) resulted in spurious predictions and poor ability to extrapolate in both the 10-fold and 50-fold cross-validation tests. This work has demonstrated that this is due to the fact that the malaria prevalence data being modelled are overdispersed relative to the binomial sampling assumption and spatial-only model. That is, the assumption is violated that the infection status of each individual in a given sample is independent of the others, given the estimated prevalence estimate at that location.

This should not be surprising from an epidemiological perspective, given that the infections in a given place do not arise independently—each infection is caused by another. This gives rise to local noise, either at the level of a pixel or group of pixels (that particular location may have some risk factor not accounted for by the smooth spatial model), or at the level of the observation (on the day of sampling, that population may have had a higher or lower than usual prevalence). The INLA-based model's behaviour in this case is an attempt to capture these small-scale variations with a very 'wiggly' spatial random effect, i.e. one with rapid decay with increasing distance. It favours this parameter configuration on overdispersed data because the observation variance is fixed when using a binomial likelihood, and the variance is not sufficiently large to explain the data. This issue of poor identifiability between the observation-level variance and the lengthscale of a Gaussian processes has previously been described ([69], see Fig. 5.4), and can be resolved in classical (and model-based) geostatistics with the use of an independent 'nugget' effect either on each observation or each observed location [1]. Despite also using a binomial observation distribution, FRK does not suffer the same pitfall because it includes a type of spatial nugget effect in its 'small-scale' effect parameter.

For malaria prevalence modelling with INLA, this analysis suggests that a more reliable 'default' model than the standard binomial observation model would be one which includes additional observation-level random noise. This can be achieved by using a Beta-binomial or Gaussian (on the observed prevalences or on the empirical-logit scale). Both of these options have an additional observation-level variance parameter that can be used to explain the overdispersion relative to the binomial. Of these, the Beta-binomial is most likely to be generally applicable to malaria prevalence data, since it is able to accurately account for observation errors in the common situation where only very few of the individuals tested are infected. It is worth noting that fitting with a Gaussian response is substantially more computationally efficient in INLA, and so may be preferable if computation time is a major constraint. An alternative approach would be to include an independent observation-level random effect in the model specification.

Whilst FRK scaled well to large datasets (generally taking slightly less time than INLA) and performed well in both the 10-fold and 50-fold extrapolation comparisons, for continental-scale modelling, specifying the model in such a way that it was both computationally scalable and avoided the spurious oscillating effect of the basis functions was not achieved using with modest model modifications. Whilst less noticeable, similar patterns are visible in the national-scale analysis in parts of North-Western and far North-Eastern Kenya where no data are available to inform such a prediction. Given these issues, significant care must be taken, when applying FRK to mapping of sparse malariometric data, to avoid these spurious predictions that are driven by computationally convenient approximations rather than data.

Comparing four methodologically different techniques has its limitations. One such limitation is that the inherent differences of the methods make a comparison somewhat difficult. For example, the likelihoods are different as well as the underlying model structure and/or covariance functions. Thus, each method has its own measures and an INLA goodness of fit measure cannot directly be compared with that of SpRF and vice versa. For this analysis, this has been mitigated by focusing on the outputs—predictive maps and cross-validation results.

Another aspect of interest is the parameter settings. There are many different parameter settings for each method. This analysis selected the commonly used (default) parameter settings and even though several different parameter settings were explored, a comprehensive exploration of the parameter space of these algorithms was not conducted. While the default parameter settings were acceptable for Kenya, it is expected that algorithms can benefit from customized parameters when running the model on the scale of Africa. The limitations of the choice of parameters is brought to light by the extent of the geographical region. Exploring optimal parameter selection is another avenue of research. Furthermore, there might be other parameter settings that can make the inference approximations of the different models more comparable.

From a practitioner's point of view, it is challenging to adopt a new method for spatial modelling mostly because it takes a long time to learn the methodology and write code to produce meaningful output. This is a significant barrier to entry. If the methods discussed provide tuning functions that explore the parameter space and select a set of parameters that enables the practitioner to build a good model, it would increase the usability of these methods.

An in-depth investigation of strengths and weaknesses of the models would be another avenue of interest. One option is to construct a meta-model that can predict the best model based on features of different locations [70]. Such a meta-model could combine the strengths of the diverse models to make a stronger prediction. The findings of this paper should be of use for those creating, interpreting or working with spatial data, as a baseline comparison of new computational geostatistical models.

Appendix

See Tables 3, 4, 5, 6 and Figs. 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22

Table 3 Cross-validation results of the four models

Fold	Model	RMSE	Correlation	% points with absolute error less than		
				0.05	0.1	0.2
10-fold	INLA	0.181	0.235	69.211	76.316	86.053
	GPBoost	0.127	0.646	52.632	73.947	**93.158**
	SpRF	0.132	0.641	**69.474**	**80.263**	91.053
	FRK	**0.125**	**0.661**	57.105	76.579	92.105
50-fold	INLA	0.124	0.683	**69.474**	**80.789**	87.895
	GPBoost	**0.11**	**0.751**	65.789	**80.789**	90
	SpRF	0.121	0.702	67.632	78.947	**90.526**
	FRK	0.123	0.702	66.053	79.211	90.263

Boldface denotes the best score in each column for each of the cross-validation experiments

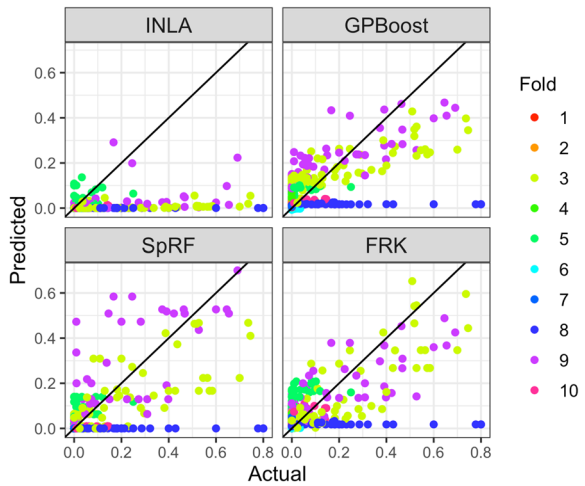

Fig. 10 Model predictions of the four models vs actual prevalences using 10-fold CV

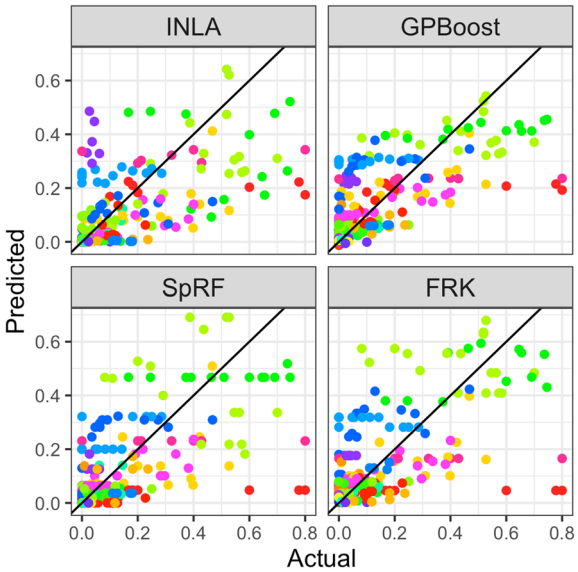

Fig. 11 Model predictions of the four models vs actual prevalences using 50-fold CV with folds in different colours

Detailed cross-validation results

Point predictions

Table 3 gives the results for 10-fold and 50-fold cross-validation. For each set of folds it gives the Root Mean Square Error (RMSE), correlation coefficient between the predicted and actual values and the percentage of observations that have an absolute error $(|\text{predicted} - \text{actual}|)$ less than a specified threshold (thresholds of 0.05, 0.1 and 0.2 considered). As noted earlier, FRK and GPBoost have the best RMSE and correlation values for 10-fold and 50-fold cross-validation respectively. SpRF gives the best performance in terms of the percentage of observations with absolute error less than 0.05 and 0.1. Compared to the other models, INLA performs poorly for the 10-fold cross-validation, with a higher RMSE and significantly lower correlation coefficient. However, it gets a high percentage of observations with absolute error less than the three thresholds. This is because a large number of observations have low prevalence values. This is further illustrated in Fig. 10, which shows the actual and predicted values using 10-fold cross-validation for each model.

Figure 10 shows the points by cross-validation fold as determined in Fig. 2a As the folds are determined by k -means clustering, observations in each fold lie close together. Data points in most folds have similar prevalence values, apart from the points in Folds 3, 8 and 9 which have a broad range of values. The points assigned to Fold 8 are difficult to predict for all four models. These points are along the coast near the city of Mombasa and are somewhat isolated from other clusters, which might be a contributing reason.

Figure 11 shows 50-fold cross-validation results for the four models while Fig. 12 shows their interval predictions. As shown in Table 3, GPBoost achieves better results in terms of RMSE and correlation. It has the

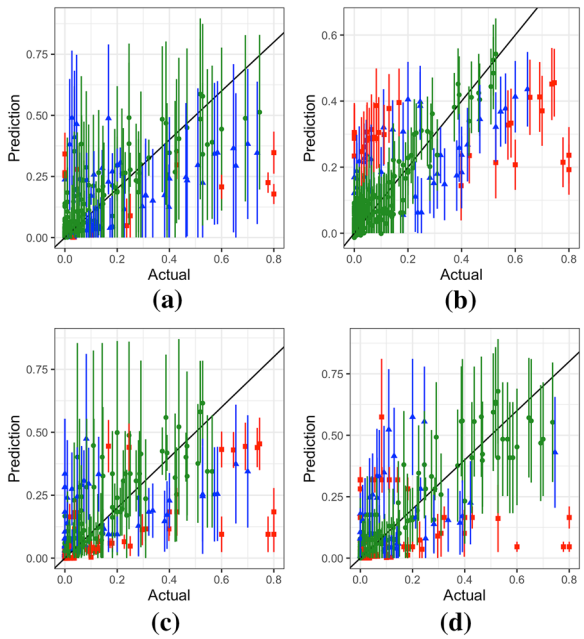

Fig. 12 Interval predictions for 50-fold cross-validation for **a** INLA, **b** GPBoost, **c** SpRF and **d** FRK

same performance as INLA for the highest percentage of observations with absolute error less than 0.1. INLA has the highest percentage of observations with absolute error less than 0.05 and SpRF has the highest percentage of observations with absolute error less than 0.2. Figure 11 shows that certain folds perform poorly. These folds match with the locations of the poorly performing folds in the 10-fold CV scenario. Another interesting observation is that while GPBoost achieves good results for both sets of folds, it performs poorly on high prevalence observations, whereas FRK and SpRF do not appear to have this limitation.

Point predictions by location density

These results are further analyzed using the density of sampled locations, i.e. do some models find it difficult to predict observations in low density regions? Fig. 13 shows the malaria prevalence and kernel density estimates of the sampled locations on two separate maps. Figure 14 shows scatter plots of prevalence and density with points coloured by the fold. Fold 8 in the 10-fold CV, which is located around the city of Mombasa, has a broad range of prevalence values while having relatively low density. This explains the reason behind the high errors for Fold 8 (Fig. 10). When the sampled points are away from each other (low density) and the prevalence values have high variation, it is challenging for the models to predict accurately.

Figure 15 shows the absolute errors of the four models with respect to density for both 10 and 50-fold cross-validation. For 10-fold CV, Fold 8 exhibits high error rates for all four models. Considering the same set of points for 50-fold CV shows that while FRK and SpRF have similar error rates in both CV experiments (maximum ≈ 0.75), INLA and GPBoost have comparatively lower error rates (maximum ≈ 0.6). Thus, a higher number of folds

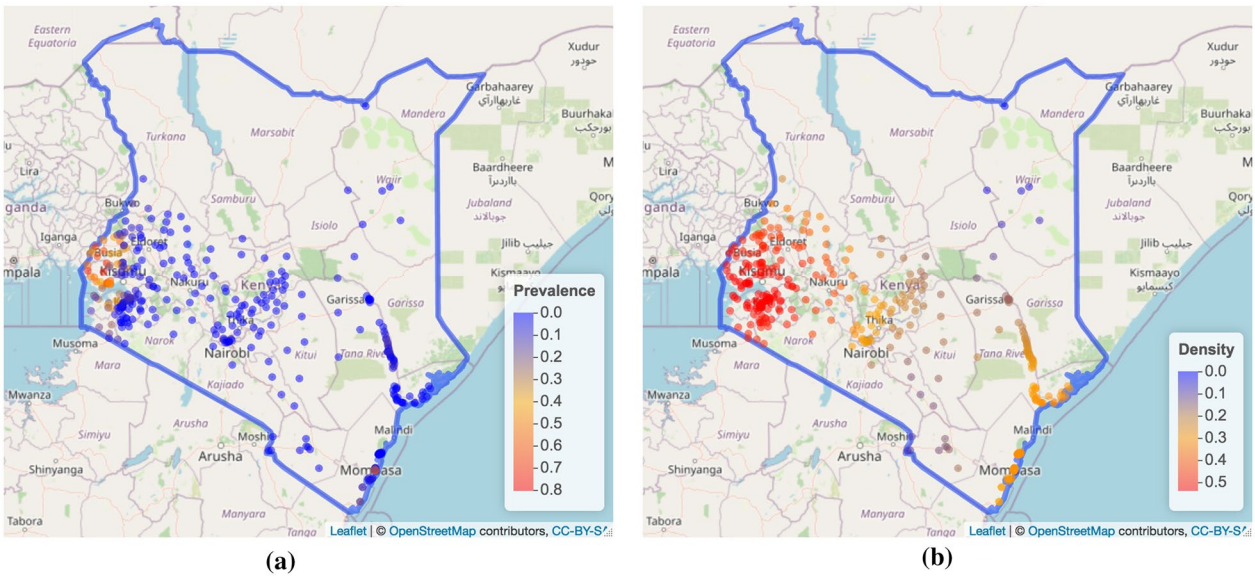

Fig. 13 **a** *P. falciparum* prevalence in Kenya for 2009 and **b** the kernel density estimates of the sampled locations

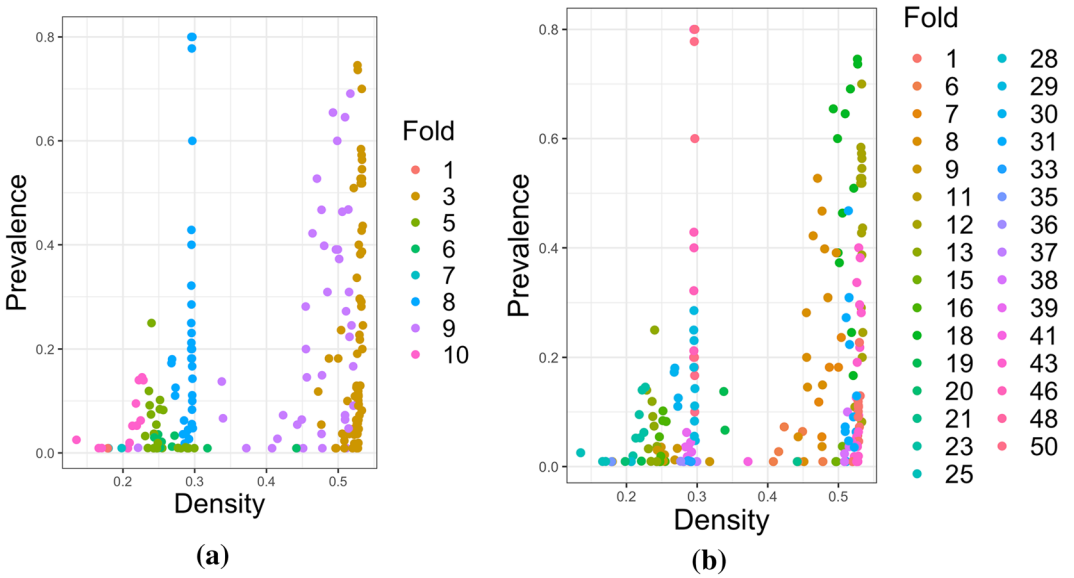

Fig. 14 *P. falciparum* prevalence and kernel density estimates of different clusters (folds) for **a** 10-fold cross-validation and **b** 50-fold cross-validation, with zero prevalence observations taken out

benefits GPBoost and INLA in this instance more than it benefits FRK or SpRF.

Moving on to the points with very low density (< 0.2), Fig. 14 shows that the prevalence values of these points are relatively low. Figure 15a shows that INLA and SpRF have lower errors for these low density points compared to FRK and GPBoost. A similar outcome can be observed for the 50-fold CV case in Fig. 15b when the density of points are less than 0.2.

As seen in Fig. 14, points in high density regions (> 0.4) have a higher variation in prevalence ranging from 0 to 0.75. The 10-fold CV results in Fig. 15a show that FRK performs best for these high density points with an error < 0.4, followed by GPBoost and SpRF. INLA performs poorly on these points for 10-fold CV. For 50-fold CV (Fig. 15b) GPBoost performs best on the high density points followed by SpRF, while both FRK and INLA perform similarly.

Table 4 gives metrics for different density groups for both 10 and 50-fold CV. With low density defined as density ≤ 0.2, medium density as $0.2 <$ density ≤ 0.4 and high density as density > 0.4, the absolute errors of all four models are small for low density points, while the percentage of observations with absolute error less than 0.05, 0.1 and 0.2 are quite high for both 10 and 50-fold cross-validation sets. In both CV sets, INLA has the lowest RMSE with SpRF following closely. For both 10 and 50-fold CV the correlation coefficients between the actual and the predicted values are negative. This indicates that most prevalence values are close to zero in low density locations. This would also explain why less points were sampled from those regions, as more points are generally sampled from high prevalence regions.

GPBoost, SpRF and FRK have higher RMSE for the medium density point set, compared to the high density point set for 10-fold CV. A similar behaviour is observed for FRK for 50-fold CV. This is due to the high absolute errors in Fold 8 as discussed previously. For the medium density points, GPBoost is preferred in terms of RMSE for both 10-fold CV and 50-fold CV. For high density points FRK is preferred for 10-fold CV while GPBoost is preferred for 50-fold. In terms of the percentage of points with absolute error less than 0.05 and 0.1, SpRF leads the other models for 10-fold CV, while INLA leads for 50-fold. For both 10 and 50-fold CV, INLA and SpRF perform better on low density points compared to the other two methods, while GPBoost and FRK perform better on high density points.

Interval predictions

As described in "Results" section, each method has a different uncertainty quantification mechanism, however this analysis has estimated the standard deviation of predictions from each model to allow comparison.

Table 5 gives the interval prediction results for all four models. Figures 5 and 12 show the interval predictions for 10 and 50-fold CV. For both 10 and 50-fold CV, SpRF has on average the smallest uncertainty intervals and

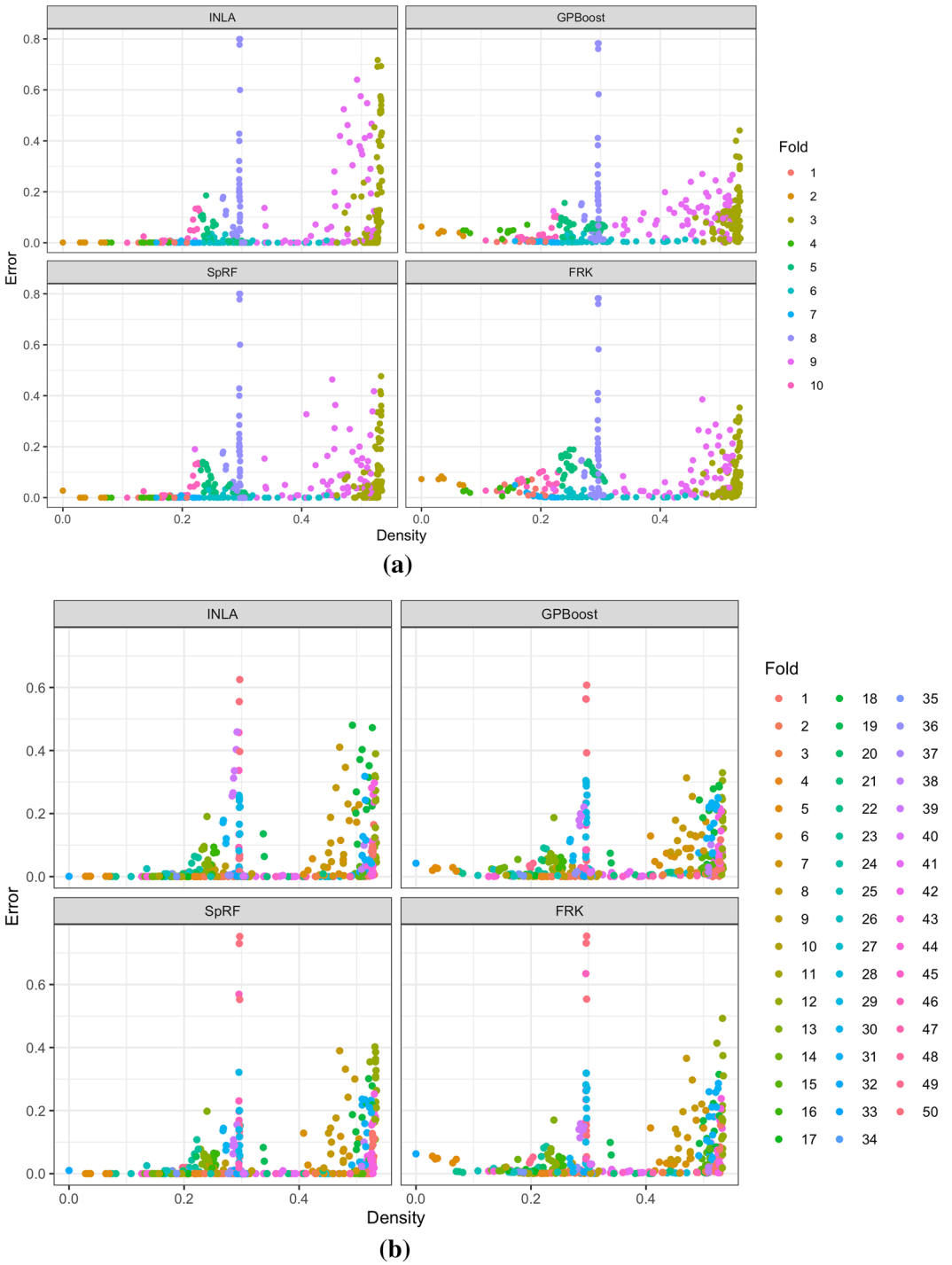

Fig. 15 Kernel density estimates of the locations and the absolute errors of the four models for **a** 10-fold cross-validation and **b** 50-fold cross-validation

the smallest number of points within one or two standard deviations of the mean. Surprisingly SpRF's interval widths are zero for 148 and 164 of the predictions for 10 and 50-fold CV respectively. Each of these points correspond to a prediction (median) of zero prevalence, and the majority correspond to an observed prevalence of zero. However, the mean value of the response at nearly all of these locations is small but non-zero, and

Table 4 Cross-validation results grouped by density of sampled locations

Fold	Density	Model	RMSE	Corr	% points with absolute error less than		
					0.05	0.1	0.2
10-fold	Low	INLA	**0.005**	− 0.234	100	100	100
		GPBoost	0.028	− 0.189	96	100	100
		SpRF	0.006	− 0.059	100	100	100
		FRK	0.048	− 0.021	62	98	100
	Medium	INLA	0.141	− 0.003	75.723	82.659	93.642
		GPBoost	**0.138**	− 0.021	60.694	85.549	94.22
		SpRF	0.143	− 0.025	72.832	80.347	93.642
		FRK	0.145	− 0.035	58.382	76.301	94.22
	High	INLA	0.24	0.305	52.229	61.783	73.248
		GPBoost	0.134	0.788	29.936	52.866	89.809
		SpRF	0.14	0.737	56.051	73.885	85.35
		FRK	**0.119**	0.828	54.14	70.064	87.261
50-fold	Low	INLA	**0.005**	− 0.128	100	100	100
		GPBoost	0.017	− 0.202	98	100	100
		SpRF	0.006	− 0.077	100	100	100
		FRK	0.021	− 0.17	96	100	100
	Medium	INLA	0.123	0.501	73.41	83.815	90.173
		GPBoost	**0.117**	0.511	73.41	84.393	91.908
		SpRF	0.121	0.435	71.676	85.549	96.532
		FRK	0.133	0.31	75.723	83.237	91.329
	High	INLA	0.143	0.776	55.414	71.338	81.529
		GPBoost	**0.119**	0.809	47.134	70.701	84.713
		SpRF	0.14	0.742	52.866	64.968	80.892
		FRK	0.129	0.784	45.86	68.153	85.987

Boldface denotes the best RMSE score for a given number of cross validation folds and point density

Table 5 Interval prediction results of the four models

Fold	Model	Mean width	Std. Dev. Width	Points within (%)	
				1SD	2SD
10-fold	INLA	0.102	0.048	75	87.105
	GPBoost	0.136	0.014	84.211	95.526
	SpRF	0.071	0.106	37.105	55
	FRK	0.1	0.072	83.421	92.368
50-fold	INLA	0.096	0.091	81.053	95.789
	GPBoost	0.112	0.014	81.053	91.053
	SpRF	0.056	0.086	28.158	42.895
	FRK	0.062	0.068	74.474	85

Mean width refers to the average of the predicted standard deviations, while Std. Dev. width refers to their standard deviation

so the prevalence at these points does not lie within any number of standard deviations of the mean, contributing to the low percentages for SpRF in Table 5.

For both 10 and 50-fold CV, GPBoost has the largest mean interval widths but smallest standard deviation, suggesting that it predicts consistently high width intervals for most observations. For 10-fold CV, this results in the highest percentage of points lying within one or two standard deviations. For 50-fold CV INLA has a higher percentage of points lying within each type of interval, which may be accounted for by the higher variation in INLA's interval widths.

Figure 16 shows the kernel density estimates of the locations and the respective interval widths of the four models for both 10 and 50-fold CV. For both 10 and 50-fold CV, GPBoost has similar widths for all observations. There is a slight increase in width for low density points. However, there is not much variation in width with respect to the density. In contrast, FRK, INLA and SpRF have varying interval widths for different folds. There is also high variation for locations with high density, mostly likely because of the variation in prevalence. For both 10 and 50-fold CV, FRK has relatively high width values for low density

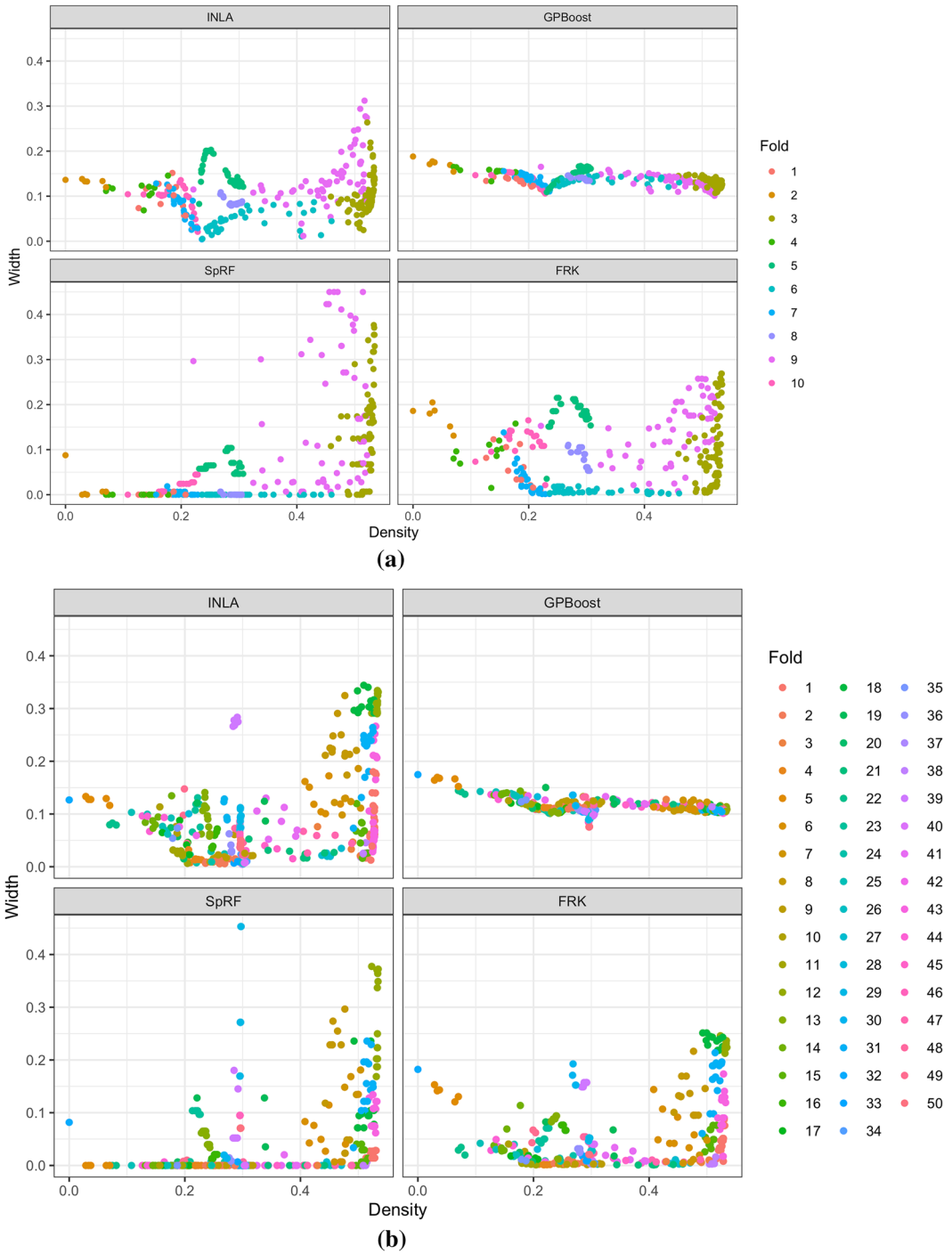

Fig. 16 Kernel density estimates of the locations and the interval widths of the four models for **a** 10-fold cross-validation and **b** 50-fold cross-validation

points. Conversely, INLA and SpRF have low width values for low density points. Similar to the point predictions, a high variation of width occurs for medium density points (density ≈ 0.3) for FRK, INLA and SpRF.

Effects of input noise on INLA

Figure 6(ai) shows INLA predicting a flat near-zero prevalence over most of Africa when trained on the observation data, a behaviour that is not replicated by fitting the model to either set of simulated data. This

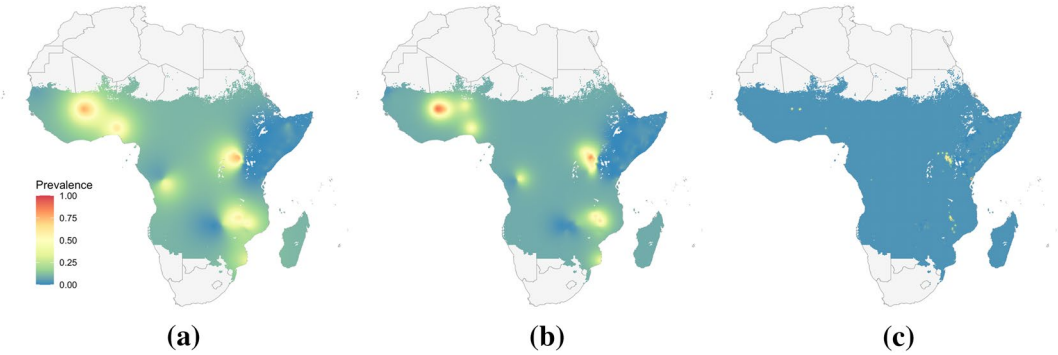

Fig. 17 Predicted prevalence from INLA when fit to simulated data at observation locations with **a** no added noise, **b** added Gaussian noise of standard deviation 0.4 and **c** added Gaussian noise of standard deviation 1.2

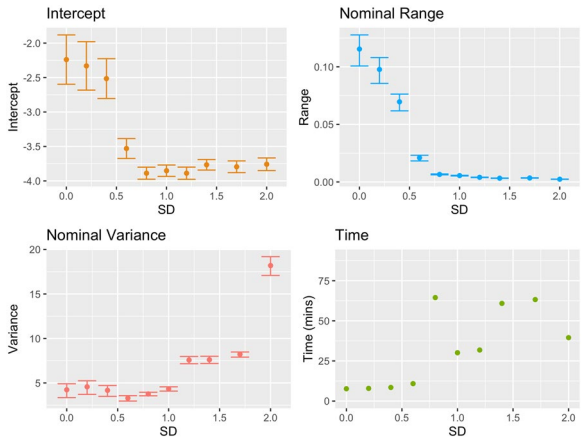

Fig. 18 Posterior means of the intercept, range and variance for the INLA-based model fit using simulated data at the observation locations with added Gaussian noise of varying standard deviation. The bottom right plot shows the time taken to fit the model to each of the datasets. Error bars show posterior interquartile ranges

behaviour may be due to the noise in the observation data, which is visible in Fig. 3a particularly in Uganda. In contrast, the simulated data at the same locations, shown in Fig. 3b, appears much smoother.

This working hypothesis is examined by adding Gaussian noise to the simulated data. Prevalence values sampled from the MAP raster in Fig. 3d with locations based on the observation points were transformed using the logit function. Gaussian white noise with chosen standard deviations was added to the transformed values, before being brought back to values between 0 and 1 via the inverse logit. Binomial samples for the number of positive tests were then drawn using these prevalences and INLA was fit to this data.

Predictions from INLA fitted to data with three different levels of noise are shown in Fig. 17. Figure 18 shows posterior means and interquartile ranges for the intercept, range, and variance of the fitted models as the standard deviation of the added noise increases, as well as the time taken to fit each model and generate predictions.

As the standard deviation of added noise increases, both the intercept and spatial range fall, and predictions become less correlated between locations. The time taken jumps for standard deviations above 0.6 and presumably the model has difficulty converging. With greater noise, the model predicts a flat prevalence away from the simulated data dependent on the value of the intercept, and its output in Fig. 17c resembles the predictions in Fig. 6(ai) of the model trained on the observation data. These

Table 6 Cross-validation results for INLA-based models using a binomial and Beta-binomial response

Fold	Model	RMSE	Correlation	% points with absolute error less than		
				0.05	0.1	0.2
10-fold	INLA binomial	0.181	0.235	69.211	76.316	86.053
	INLA Beta-binomial	0.166	0.457	70.263	76.316	85.789
50-fold	INLA binomial	0.124	0.683	69.474	80.789	87.895
	INLA Beta-binomial	0.115	0.74	71.053	80.263	89.474

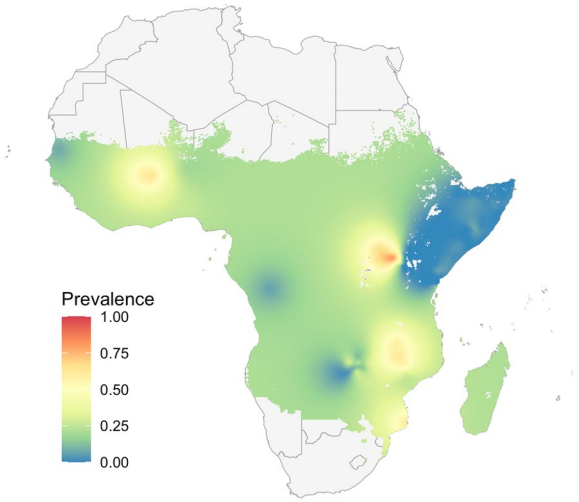

Fig. 19 Predictions from an INLA-based model with a Gaussian response fit to the observation data. Values have been clipped to lie within [0, 1]

results suggest the presence of overdispersion, and that the model implemented with INLA may be misspecified. Indeed the model in Eq. (4) does not contain an independent error term. Methods to address this include adding an observational random effect to the model, or using a Beta-binomial response.

Table 6 shows the cross-validation results using the Kenya dataset for an INLA-based model using a Beta-binomial response compared to the earlier model which used a binomial response. For both 10 and 50-fold cross-validation, the Beta-binomial response leads to improved performance in terms of the root mean square error and the test correlation. For 10-fold cross-validation, the Beta-binomial model still performs worse than the other three models in Table 3, however for 50-fold cross-validation it outperforms all but GPBoost.

INLA with a Gaussian response

Due to the overdispersion when fitting the INLA-based model to the observation data, an additional model using "INLA with a Gaussian response" was tested. Predictions from this model are shown in Fig. 19 and the absence of flat predictions suggests that this response is able to resolve the overdispersion.

Prediction uncertainty

Figure 20 shows the prediction uncertainties corresponding to each of the prevalence maps over Africa in Fig. 6.

GPBoost with the Vecchia approximation

As it uses a full Gaussian process, it is unsurprising that the GPBoost model shows the least favourable computational time for larger datasets. To improve efficiency, a Vecchia approximation is available in the software, which approximates the distribution of the response as

$$p(\mathbf{y}|F(X),\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|y_{i-1},\ldots,y_1,F(X),\boldsymbol{\theta})$$

$$\approx \prod_{i=1}^{N} p(y_i|y_{N(i)},F(X),\boldsymbol{\theta}),$$

as per [46]. Here $y_{N(i)}$ is the subset of $\{y_1,\ldots,y_{i-1}\}$ containing the m_v nearest neighbours to y_i, where 'nearest neighbours' are determined by the distances between the responses' corresponding locations. The parameter m_v determines the number of neighbours to use during fitting, while a separate parameter, $m_{v,p}$, controls the number of neighbours used for prediction. The above approximation of the density results in a sparse Cholesky factor of the precision matrix, with the number of selected neighbours impacting this sparsity and the accuracy of the approximation [58]. The approximation additionally requires a choice of ordering of the observed responses $\{y_1,\ldots,y_n\}$, which by default is taken to be a random ordering of the data.

Figure 21 shows GPBoost's predictions when using a Vecchia approximation with several values of m_v and $m_{v,p}$. Uncertainty predictions were not produced as they are not currently well supported in the software when using the Vecchia approximation.

Applying the Vecchia approximation introduces artifacts to GPBoost's predictions. Figure 21a shows sharp discontinuities, which were prominent whenever low values of m_v and $m_{v,p}$ were used. Experiments using the Kenya data suggested that the discontinuities could be prevented by increasing m_v and $m_{v,p}$, however for the dataset on the continent scale there was a computational cost for doing so. Increasing $m_{v,p}$ from 30 to 150 while keeping m_v fixed at 30 had a relatively small impact on the computation time, which surprisingly decreased from 11.09 min to 9.56 min, but the required memory jumped from 425 MB to 20,593 MB. Meanwhile, increasing both m_v and $m_{v,p}$ to 150 greatly increased the computation time, requiring over 3.6 h to run, much longer than when the Vecchia approximation was not applied. Additionally, this model configuration required 20,722 MB of RAM. These examples suggest that increasing m_v primarily increases the computation time required without

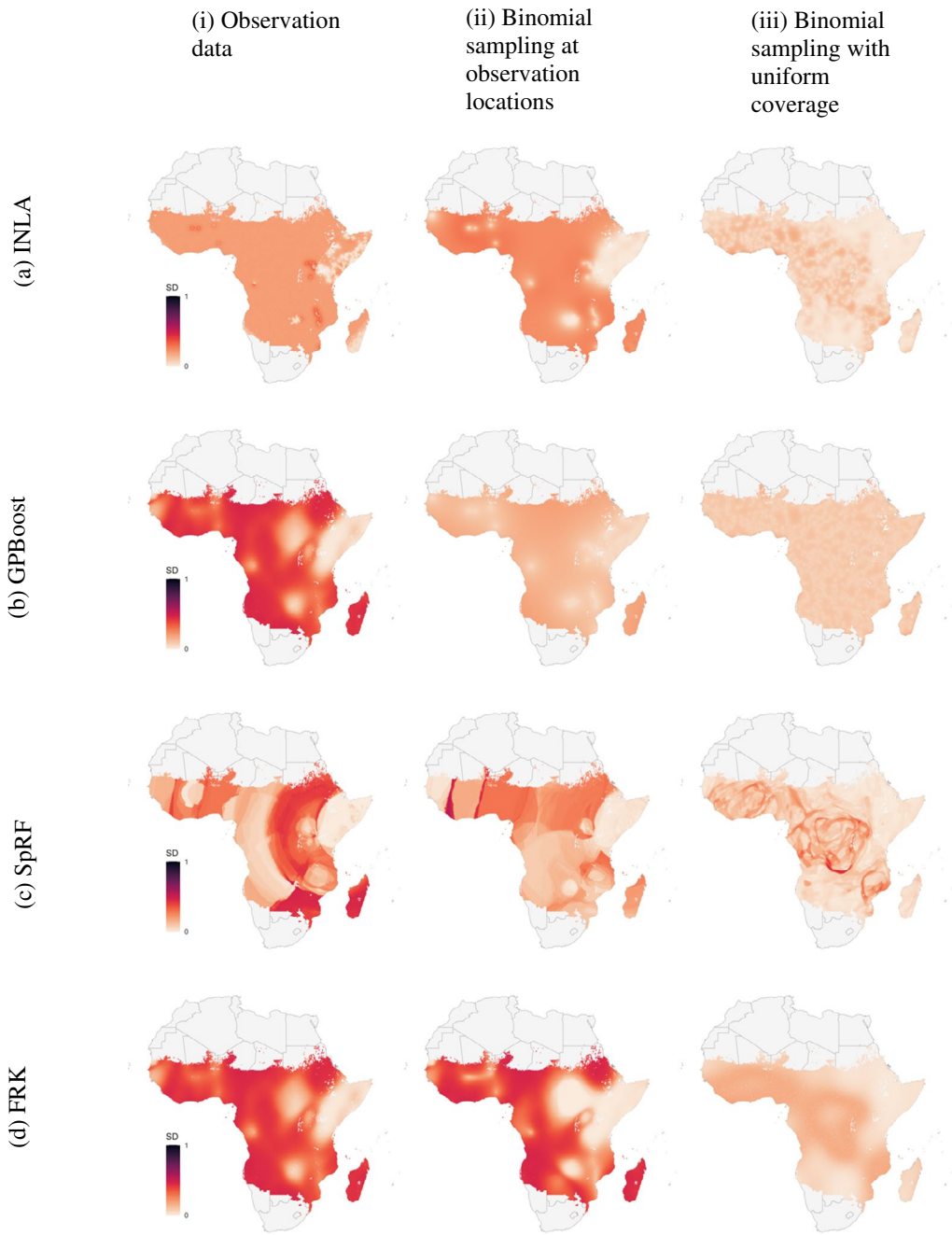

Fig. 20 Predicted standard deviations for each of the maps shown in Fig. 6

affecting the RAM usage, while increasing $m_{v,p}$ increases the required RAM, with a smaller impact on computation time. Despite the increased computational requirements, neither adjustment to the parameters completely removed the discontinuities.

One benefit of using the Vecchia approximation is an improvement in scaling behaviour, even if the computational requirements on an individual dataset depend strongly on the choice of m_v and $m_{v,p}$. Figure 22 shows the computational results from Fig. 7, with an

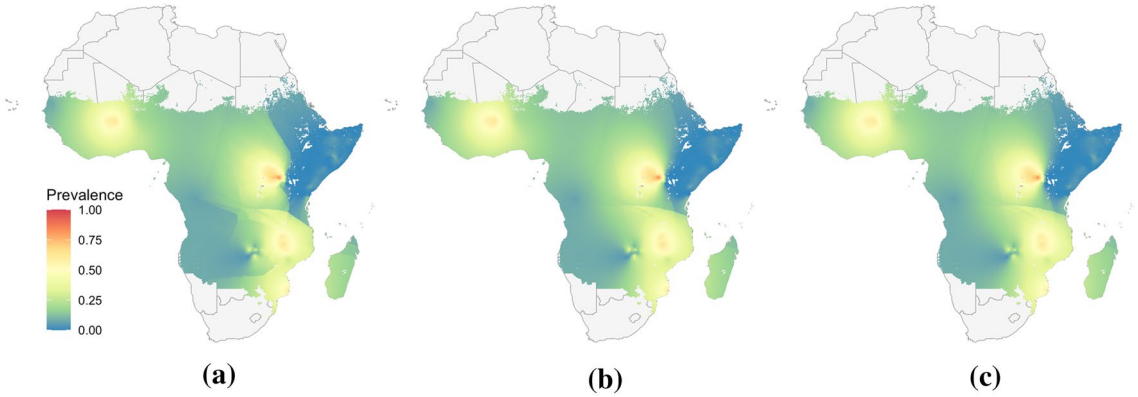

Fig. 21 *P. falciparum* prevalence predictions for GPBoost when using the Vecchia approximation for various values of the nearest neighbour parameters, m_v and $m_{v,p}$. **a** uses $m_v = m_{v,p} = 30$, **b** uses $m_v = 30$ and $m_{v,p} = 150$, while **c** uses $m_v = m_{v,p} = 150$

additional plot for GPBoost using the Vecchia approximation. Parameters m_v and $m_{v,p}$ were held fixed at 30 and 150 respectively, and the model shows favourable computation times compared to the full GP version of the model and SpRF, though is slower than INLA and FRK. These results highlight an obstacle to applying GPBoost to large scale data. Using the full Gaussian process can result in large computation times, while applying the Vecchia approximation introduces artifacts which require sacrifices in computational efficiency to remove.

GPBoost's computation times are amplified by the high value of the `nrounds` parameter, which has been set to 247 following available tutorials. As described in "GPBoost-based model", this parameter controls the number of optimization steps during fitting. When fit to the malaria datasets used throughout this paper, the log-likelihood generally stopped increasing after 5 to 10 steps, suggesting that 247 training steps is unnecessarily high for the considered data and model. Figure 22 additionally includes the times taken for GPBoost with `nrounds` reduced to 5, and the scaling behaviour is greatly improved by this change, with the fastest computation time among all of the implemented models. Additional experimentation however found that reducing the number of rounds had little effect on the high RAM requirements for large values of $m_{v,p}$; something which may be necessary to minimize the discontinuities and noise in the predictions. Reducing `nrounds` would also improve the efficiency of the GPBoost model when no Vecchia approximation is used, however would not change the overall scaling behaviour.

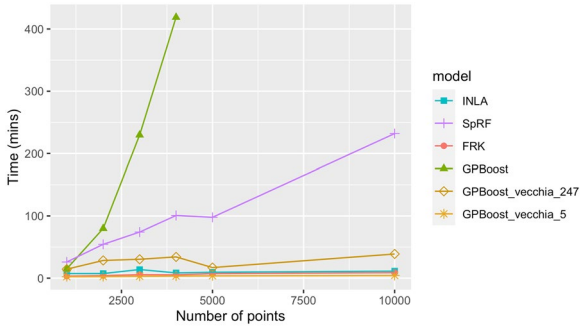

Fig. 22 Time taken for GPBoost with the Vecchia approximation using values of `nrounds` = 247 and `nrounds` = 5 applied for simulated datasets of various sizes, compared to the times in Fig. 7

Acknowledgements
The authors would like to thank Håvard Rue, Andrew Zammit-Mangion, Matthew Sainsbury-Dale, Fabio Sigrist and Noel Cressie for their correspondence and help with setting up and troubleshooting models. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. J.A. Flegg's research is supported by the Australian Research Council (DP200100747, FT210100034). J.A. Flegg and N. Golding's research is supported by the National Health and Medical Research Council (APP2019093).

Author contributions
SK, NG, and JF conceptualized the study. NG and JF curated data. Formal analysis, software development, and visualization were carried out by SW and SK. Methodology was determined by all authors. The manuscript was written and edited by all authors.

Funding
J.A. Flegg's research is supported by the Australian Research Council (DP200100747, FT210100034). J.A. Flegg and N. Golding's research is supported by the National Health and Medical Research Council (APP2019093).

Availability of data and materials
All model output and programming scripts for this work are available at https://github.com/sevvandi/ supplementary_material/tree/master/stcompare.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 May 2023 Accepted: 18 October 2023
Published online: 21 November 2023

References

1. Diggle P, Ribeiro Jr PJ. Model-based geostatistics. Springer; 2007.
2. Martínez-Minaya J, Cameletti M, Conesa D, Pennino MG. Species distribution modelling: a statistical review with focus in spatio-temporal issues. Stoch Environ Res Risk Assess. 2018;32:3227–44.
3. Holdaway MR. Spatial modelling and interpolation of monthly temperature using kriging. Clim Res. 1996;6(3):215–25.
4. Samalot A, Astitha M, Yang J, Galanis G. Combined Kalman filter and universal kriging to improve storm wind speed predictions for the northeastern United States. Weather Forecast. 2019;34(3):587–601.
5. Mulla D. Mapping and managing spatial patterns in soil fertility and crop yield. In: Proceedings of soil specific crop management: a workshop on research and development issues. Wiley Online Library; 1993. pp. 15–26.
6. Kuntz M, Helbich M. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. Int J Geogr Inf Sci. 2014;28(9):1904–21.
7. Lai YS, Zhou XN, Utzinger J, Vounatsou P. Bayesian geostatistical modelling of soil-transmitted helminth survey data in the People's Republic of China. Parasit Vectors. 2013;6(1):359.
8. Oliver M, Muir K, Webster R, Parkes S, Cameron A, Stevens M, et al. A geostatistical approach to the analysis of pattern in rare disease. J Public Health. 1992;14(3):280–9.
9. Scholte RG, Gosoniu L, Malone JB, Chammartin F, Utzinger J, Vounatsou P. Predictive risk mapping of schistosomiasis in Brazil using Bayesian geostatistical models. Acta Trop. 2014;132:57–63.
10. Nicholson MC, Mather TN. Methods for evaluating Lyme disease risks using geographic information systems and geospatial analysis. J Med Entomol. 1996;33(5):711–20.
11. Alimi TO, Fuller DO, Quinones ML, Xue RD, Herrera SV, Arevalo-Herrera M, et al. Prospects and recommendations for risk mapping to improve strategies for effective malaria vector control interventions in Latin America. Malar J. 2015;14(1):519.
12. Omumbo JA, Noor AM, Fall IS, Snow RW. How well are malaria maps used to design and finance malaria control in Africa? PLoS ONE. 2013;8(1): e53198.
13. Weiss DJ, Lucas TC, Nguyen M, Nandi AK, Bisanzio D, Battle KE, et al. Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study. Lancet. 2019;394(10195):322–31.
14. Chipeta MG, Giorgi E, Mategula D, Macharia PM, Ligomba C, Munyenyembe A, et al. Geostatistical analysis of Malawi's changing malaria transmission from 2010 to 2017. Wellcome Open Res. 2019;4:57.
15. Noor AM, Kinyoki DK, Mundia CW, Kabaria CW, Mutua JW, Alegana VA, et al. The changing risk of Plasmodium falciparum malaria infection in Africa: 2000–10: a spatial and temporal analysis of transmission intensity. Lancet. 2014;383(9930):1739–47.
16. Ashton RA, Kefyalew T, Rand A, Sime H, Assefa A, Mekasha A, et al. Geostatistical modeling of malaria endemicity using serological indicators of exposure collected through school surveys. Am J Trop Med Hyg. 2015;93(1):168–77.
17. Gething PW, Casey DC, Weiss DJ, Bisanzio D, Bhatt S, Cameron E, et al. Mapping Plasmodium falciparum mortality in Africa between 1990 and 2015. N Engl J Med. 2016;375(25):2435–45.
18. Bertozzi-Villa A, Bever CA, Koenker H, Weiss DJ, Vargas-Ruiz C, Nandi AK, et al. Maps and metrics of insecticide-treated net access, use, and nets-per-capita in Africa from 2000–2020. Nat Commun. 2021;12(1):3589.
19. Flegg JA, Patil AP, Venkatesan M, Roper C, Naidoo I, Hay SI, et al. Spatiotemporal mathematical modelling of mutations of the dhps gene in African Plasmodium falciparum. Malar J. 2013;12(1):249.
20. Flegg JA, Humphreys GS, Montanez B, Strickland T, Jacome-Meza ZJ, Barnes KI, et al. Spatiotemporal spread of Plasmodium falciparum mutations for resistance to sulfadoxine-pyrimethamine across Africa, 1990–2020. PLoS Comput Biol. 2022;18(8): e1010317.
21. Amoah B, Giorgi E, Heyes DJ, van Burren S, Diggle PJ. Geostatistical modelling of the association between malaria and child growth in Africa. Int J Health Geogr. 2018;17(1):7.
22. Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat Commun. 2010;1(1):104.
23. Hay SI, Snow RW. The Malaria atlas project: developing global maps of malaria risk. PLoS Med. 2006;3(12): e473.
24. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IRF, Johnston GL, et al. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. Malar J. 2011;10:378.
25. Bhatt S, Cameron E, Flaxman SR, Weiss DJ, Smith DL, Gething PW. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. J R Soc Interface. 2017;14(134):20170520.
26. Hensman J, Fusi N, Lawrence ND. Gaussian processes for big data. arXiv preprint arXiv:13096835. 2013.
27. Rasmussen CE, Nickisch H. Gaussian processes for machine learning (GPML) toolbox. J Mach Learn Res. 2010;11:3011–5.
28. Park C, Apley D. Patchwork kriging for large-scale Gaussian process regression. J Mach Learn Res. 2018;19:1–43.
29. Quiñonero-Candela J, Rasmussen CE. A unifying view of sparse approximate Gaussian process regression. J Mach Learn Res. 2005;6:1939–59.
30. Datta A, Banerjee S, Finley AO, Gelfand AE. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J Am Stat Assoc. 2016;111(514):800–12.
31. Moraga P, Dean C, Inoue J, Morawiecki P, Noureen SR, Wang F. Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: a case study predicting malaria risk in Mozambique. Spat Spatiotemporal Epidemiol. 2021;39: 100440.
32. Pfeffer DA, Lucas TCD, May D, Harris J, Rozier J, Twohig KA, et al. MalariaAtlas: an R interface to global malariometric data hosted by the Malaria Atlas Project. Malar J. 2018;17(1):352.
33. Odhiambo JN, Kalinda C, Macharia PM, Snow RW, Sartorius B. Spatial and spatio-temporal methods for mapping malaria risk: a systematic review. BMJ Glob Health. 2020;5(10): e002919.
34. Adigun AB, Gajere EN, Oresanya O, Vounatsou P. Malaria risk in Nigeria: Bayesian geostatistical modelling of 2010 malaria indicator survey data. Malar J. 2015;14:156.
35. Kazembe LN, Kleinschmidt I, Holtz TH, Sharp BL. Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data. Int J Health Geogr. 2006;5:41.
36. Nzabakiriraho JD, Gayawan E. Geostatistical modeling of malaria prevalence among under-five children in Rwanda. BMC Public Health. 2021;21:369.
37. Bhatt S, Weiss D, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526(7572):207–11.
38. Kabaria CW, Molteni F, Mandike R, Chacky F, Noor AM, Snow RW, et al. Mapping intra-urban malaria risk using high resolution satellite imagery: a case study of Dar es Salaam. Int J Health Geogr. 2016;15(1):26.
39. Kapwata T, Gebreslasie MT. Random forest variable selection in spatial malaria transmission modelling in Mpumalanga Province, South Africa. Geospat Health. 2016;11(3):434.
40. Zammit-Mangion A, Cressie N, Shumack C. On statistical approaches to generate level 3 products from satellite remote sensing retrievals. Remote Sens (Basel). 2018;10(1):155.
41. Sakizadeh M, Zhang C. Health risk assessment of nitrate using a probabilistic approach in groundwater resources of western part of Iran. Environ Earth Sci. 2020;79(1):43.
42. Wu J, Jia P, Feng T, Li H, Kuang H. Spatiotemporal analysis of built environment restrained traffic carbon emissions and policy implications. Transp Res D Transp Environ. 2023;121:103839.
43. STcompare code and datasets. https://github.com/sevvandi/supplementary_material/tree/master/stcompare. Accessed May 2023.
44. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Series B Stat Methodol. 2009;71(2):319–92.
45. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial

differential equation approach. J R Stat Soc Series B Stat Methodol. 2011;73(4):423–98.
46. Gaussian SF, Boosting P. J Mach Learn Res. 2022;23(232):1–46.
47. Hengl T, Nussbaum M, Wright MN, Heuvelink GB, Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 2018;6: e5518.
48. Zammit-Mangion A, Cressie N. FRK: an R package for spatial and spatio-temporal prediction with large datasets. J Stat Softw. 2021;98(4):1–48.
49. Sadoine ML, Smargiassi A, Ridde V, Tusting LS, Zinszer K. The associations between malaria, interventions, and the environment: a systematic review and meta-analysis. Malar J. 2018;17(1):73.
50. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian computing with INLA: a review. Annu Rev Stat Appl. 2017;4(1):395–421.
51. R-INLA Project. https://www.r-inla.org/home. Accessed Oct 2021.
52. Bakka H, Rue H, Fuglstad GA, Riebler A, Bolin D, Illian J, et al. Spatial modeling with R-INLA: a review. Wiley Interdiscip Rev Comput Stat. 2018;10(6): e1443.
53. Gómez-Rubio V. Bayesian inference with INLA. Boca Raton: Chapman & Hall/CRC Press; 2020.
54. Wang X, Ryan YY, Faraway J. Bayesian regression modeling with INLA. Boca Raton: Chapman & Hall/CRC Press; 2018.
55. Kang SY, Battle KE, Gibson HS, Ratsimbasoa A, Randrianarivelojosia M, Ramboarina S, et al. Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to assess *Plasmodium falciparum* endemicity trends between 2011 and 2016. BMC Med. 2018;16(1):71.
56. Moraga P. Geospatial health data: Modeling and visualization with R-INLA and shiny. Chapman & Hall/CRC Biostatistics Series; 2019.
57. Lindgren F, Rue H. Bayesian spatial modelling with R-INLA. J Stat Softw. 2015;63:1–25.
58. Kang M, Katzfuss M. Correlation-based sparse inverse Cholesky factorization for fast Gaussian-process inference. Stat Comput. 2023;33(3):56.
59. Sigrist F. Latent Gaussian model boosting. IEEE Trans Pattern Anal Mach Intell. 2022;45(2):1894–905.
60. Sigrist F. GPBoost. GitHub; 2020. Github repository, https://github.com/fabsig/GPBoost. Accessed Oct 2021.
61. Quantile MN, Forests R. Quantile Regression Forests. J Mach Learn Res. 2006;7:983–99.
62. Hengl T, Nussbaum M, Wright MN. GeoMLA. GitHub; 2021. Github repository, https://github.com/thengl/GeoMLA. Accessed Sep 2021.
63. Sainsbury-Dale M, Zammit-Mangion A, Cressie N. Modelling Big, Heterogeneous, Non-Gaussian Spatial and Spatio-Temporal Data using FRK. arXiv preprint arXiv:211002507. 2021.
64. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29.
65. Likas A, Vlassis N, Verbeek J. The global k-means clustering algorithm. Pattern Recognit. 2003;36(2):451–61.
66. Humphreys JM, Elsner JB, Jagger TH, Pau S. A Bayesian geostatistical approach to modeling global distributions of Lygodium microphyllum under projected climate warming. Ecol Modell. 2017;363:192–206.
67. Cressie N. Statistics for spatial data. John Wiley & Sons; 2015
68. Zammit-Mangion A, Sainsbury-Dale M. Package 'FRK'; 2023. Package documentation, https://cran.r-project.org/web/packages/FRK/FRK.pdf. Accessed Apr 2023.
69. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. vol. 14. The MIT Press; 2006
70. Wang X, Smith-Miles K, Hyndman R. Rule induction for forecasting method selection: meta-learning the characteristics of univariate time series. Neurocomputing. 2009;72(10–12):2581–94.

Publisher's Note