INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Foo YS, Flegg JA. 2024 A spatio-temporal model of multi-marker antimalarial resistance. *J. R. Soc. Interface* **21**: 20230570.
https://doi.org/10.1098/rsif.2023.0570

Received: 1 October 2023
Accepted: 12 December 2023

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
bioinformatics

Keywords:
antimalarial drug resistance, spatio-termporal mapping, sulfadoxine-pyrimethamine, haplotype frequency estimation, latent multinomial

Author for correspondence:
Yong See Foo
e-mail: yongsee.foo@unimelb.edu.au

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6992018.

THE ROYAL SOCIETY PUBLISHING

A spatio-temporal model of multi-marker antimalarial resistance

Yong See Foo and Jennifer A. Flegg

School of Mathematics and Statistics, The University of Melbourne, Parkville, Australia

YSF, 0000-0003-3010-9106; JAF, 0000-0002-8809-726X

The emergence and spread of drug-resistant *Plasmodium falciparum* parasites have hindered efforts to eliminate malaria. Monitoring the spread of drug resistance is vital, as drug resistance can lead to widespread treatment failure. We develop a Bayesian model to produce spatio-temporal maps that depict the spread of drug resistance, and apply our methods for the antimalarial sulfadoxine-pyrimethamine. We infer from genetic count data the prevalences over space and time of various malaria parasite haplotypes associated with drug resistance. Previous work has focused on inferring the prevalence of individual molecular markers. In reality, combinations of mutations at multiple markers confer varying degrees of drug resistance to the parasite, indicating that multiple markers should be modelled together. However, the reporting of genetic count data is often inconsistent as some studies report haplotype counts, whereas some studies report mutation counts of individual markers separately. In response, we introduce a latent multinomial Gaussian process model to handle partially reported spatio-temporal count data. As drug-resistant mutations are often used as a proxy for treatment efficacy, point estimates from our spatio-temporal maps can help inform antimalarial drug policies, whereas the uncertainties from our maps can help with optimizing sampling strategies for future monitoring of drug resistance.

1. Introduction

Malaria is a deadly disease caused by parasites that are transmitted by mosquitoes. During the treatment of a malaria infection, the parasites undergo selective pressure, favouring the survival of parasites that have genetic mutations which confer on them resistance against the drug treatment. For *Plasmodium falciparum*, the most common species of malaria parasites, drug resistance is a major threat to the control of the disease. It is therefore important to be able to quantify changes in antimalarial drug resistance, including for sulfadoxine-pyrimethamine (SP) that is used for intermittent preventive treatment for high-risk groups, namely pregnant people and young children.

Spatio-temporal trends in antimalarial drug resistance can be monitored using molecular markers known to be associated with drug resistance as a proxy of clinical efficacy. *Plasmodium falciparum* resistance against SP is characterized by mutations on the *dhps* and *dhfr* genes [1]. In fact, SP-resistant parasites often carry multiple SP-resistant mutations (a set of molecular markers or a haplotype). Genetic studies are easier to conduct and are a fraction of the cost of a clinical study—allowing for larger numbers of samples to be collected across more spatio-temporal locations [2]. This makes data from genetic studies readily amenable to model-based geostatistics.

To our knowledge, all works to date that have statistically mapped the geospatial distribution of the *dhps* and *dhfr* markers have modelled each marker separately [3–6]. Most relevant to the work presented in our paper, Flegg *et al.* [3] developed a predictive model for the geographical and temporal trends across Africa of the prevalence of mutations on the *dhps* gene of the parasite that are associated with SP resistance. A separate model was used for each

marker (*dhps* A437G, *dhps* K540E and *dhps* A581G), which models the count data with binomial distributions. Correlation between binomial probabilities are specified according to the spatio-temporal distance between their corresponding sites. Specifically, the logit transformation of the binomial probabilities are set to follow a Gaussian process (GP) distribution.

Although existing spatio-temporal mapping work has focused on individual marker mutations, molecular studies have shown that it is the presence of the double mutant haplotype *dhps* A437G/K540E and triple mutant haplotype *dhps* A437G/K540E/A581G in the parasite that are most strongly associated with an increased risk of SP treatment failure [7,8], and thus the most clinically relevant. Therefore, new modelling approaches are needed to obtain spatially continuous maps of haplotype prevalences, not just of individual marker mutation prevalences. However, not all studies report the presence or absence of each mutation simultaneously, i.e. the counts of full haplotypes are not reported. Instead, studies may only report the number of samples that carry each individual marker mutation. Since the samples corresponding to each reported count may overlap, we cannot use a multinomial distribution directly. Moreover, some studies only report on a smaller subset of all mutations of interest. We handle these discrepancies caused by partially reported data under a *latent multinomial model* [9,10], where the observed counts are treated as binary combinations of unobserved multinomial counts. In this paper, we extend the spatio-temporal GP models of individual *dhps* markers [3,6] to model the prevalences of multiple haplotypes by using a latent multinomial distribution with GPs. Handling all haplotypes within one model allows us to leverage all available data to greater utility.

This paper is structured as follows. In §2, we present the latent multinomial GP model for mapping SP drug resistance based on mutations on the *dhps* gene. We do not account for mutations on the *dhfr* gene as mutations on the *dhps* gene are more closely related to clinical SP failure and there is a triple *dhfr* mutation widely spread across Africa already [11]. This is followed by the outputs of the model in §3, showing the prevalence of each haplotype of interest over space and time. Finally, in §4, we discuss the implications of our findings.

2. Methods

In this paper, we construct a Bayesian hierarchical model that is capable of modelling partially reported multinomial count data for spatio-temporal changes in drug resistance. This modelling framework is needed to handle reporting inconsistencies found across studies on the prevalences of drug-resistant haplotypes, where different studies may report on different combinations of mutations. For example, the list of *dhps* mutations compiled by the Worldwide Antimalarial Resistance Network [12] includes: *dhps* 437G, *dhps* 540E, *dhps* 581G, *dhps* 437G-540E, *dhps* 437G-540E-581G and *dhps* 437G-540E-A581. These inconsistencies in data collection and study design significantly complicate model construction and parameter inference. We develop a latent multinomial model (§2.1) with a GP prior specification (§2.2). We describe how spatio-temporal maps of haplotype prevalences are produced under a Bayesian framework in §2.3.

2.1. Latent multinomial model formulation

Suppose that we have G molecular markers of interest for monitoring drug resistance (table 1). For our application, we have $G = 3$ molecular markers, namely *dhps* 437, *dhps* 540 and *dhps* 581. Define a *full haplotype* to be a binary string of length G recording whether each of these mutations are present (1) or absent (0). This results in $H = 2^G$ possible full haplotypes.

Table 1. Summary of notation used in §2.1.

notation	description
G	no. molecular markers under consideration
H	no. full haplotypes under consideration
N	no. data points
R_i	no. realized haplotypes that data point i reports
$\mathbf{p}_i$	vector (length H) of full haplotype probabilities at data point i
$\mathbf{z}_i$	vector (length H) of full haplotype counts at data point i
$\mathbf{y}_i$	vector (length R_i) of realized haplotype counts at data point i
$\mathbf{A}_i$	binary matrix (size $R_i \times H$) linking realized haplotypes to full haplotypes

Over space and time we have N studies (figure 1); study i ($i = 1, \ldots, N$) is conducted to estimate the prevalence $\mathbf{p}_i$ at this location of the H haplotypes associated with drug resistance. We define $\mathbf{z}_i$ to be a vector of length H that stores the number of each of the H full haplotypes for the ith study, and $n_i = z_{i,1} + z_{i,2} + \cdots + z_{i,H}$ to be the known sample size of the study. The counts $\{\mathbf{z}_i\}_{i=1}^N$ are considered unobserved (latent), each following a multinomial distribution:

$$\mathbf{z}_i \mid \mathbf{p}_i \sim \text{Mult}(n_i, \mathbf{p}_i) \quad \text{for } i = 1, \ldots, N. \tag{2.1}$$

For each $i = 1, \ldots, N$, we observe binary combinations of the latent counts, depending on the way counts are reported in the ith study. We refer to the set of full haplotypes that are included by a single observed count as a *realized haplotype*. For example, the realized haplotype 437G-540E includes the full haplotypes 110 (437G-540E-A581) and 111 (437G-540E-581G). The collection of realized haplotypes reported may vary across studies, which we encode within a *configuration matrix*. As an example, for a study which reports on the realized haplotypes *dhps* 437G, *dhps* 540E, *dhps* 581G, *dhps* 437G-540E, *dhps* 437G-540E-A581 and *dhps* 437G-540E-581G, we have the following configuration matrix:

$$\begin{array}{r} \\ \textit{dhps}\ 437\text{G} \\ \textit{dhps}\ 540\text{E} \\ \textit{dhps}\ 581\text{G} \\ \textit{dhps}\ 437\text{G-}540\text{E} \\ \textit{dhps}\ 437\text{G-}540\text{E-}581\text{G} \\ \textit{dhps}\ 437\text{G-}540\text{E-A}581 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}. \tag{2.2}$$

In this way, for an observed count of the samples with the *dhps* 437G mutation, we need to sum the number of the four haplotypes {100, 101, 110, 111}, while for an observed count of the samples with dhps 437G-540E, we only need to sum the haplotypes {110, 111}. The configuration matrix is constructed based on which haplotypes are included in each count reported by a study.

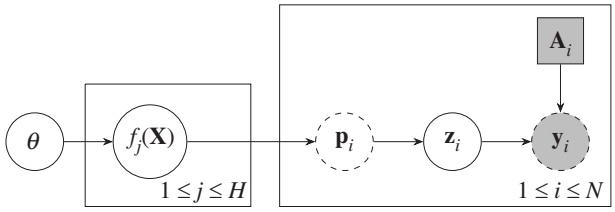

Figure 1. Graphical model for latent multinomial data with multiple populations whose haplotype prevalences $\{\mathbf{p}_i\}_{i=1}^N$ are correlated through Gaussian processes. $f_j(\mathbf{X})$ denotes the vector $(f_j(\mathbf{x}_1), \ldots, f_j(\mathbf{x}_N))$. Circles correspond to random variables while squares correspond to constant values. A shaded node indicates that the variable is observed. A dotted outline indicates that the variable is deterministically calculated from its parent variables. Variables contained within a plate (box) are replicated according to the index at the bottom right.

We denote the number of realized haplotypes reported in study i by R_i and the reported counts of the realized haplotypes by vector $\mathbf{y}_i$ of length R_i, for $i = 1, \ldots, N$. Let $\mathbf{A}_i$ be the constructed $R_i \times H$ configuration matrix of 0's and 1's, determined from the realized haplotypes reported in study i. For each $i = 1, \ldots, N$ (figure 1), the latent counts $\mathbf{z}_i$ must be a non-negative integer solution to the system

$$\mathbf{y}_i = \mathbf{A}_i \mathbf{z}_i \tag{2.3}$$

and the sample size constraint $n_i = z_{i,1} + z_{i,2} + \cdots + z_{i,H}$. Since the sample size, n_i, is known, we let the first row of $\mathbf{A}_i$ always be a vector of ones and the first observed count be the sample size, i.e. $y_{i,1} = n_i$ for all i.

To calculate the likelihood, we marginalize out the latent counts $\{\mathbf{z}_i\}_{i=1}^N$ exactly by enumerating all possible latent counts $\mathbf{z}_i$ that satisfy (2.3). The full likelihood is

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_N | \mathbf{p}_1, \ldots, \mathbf{p}_N) = \prod_{i=1}^N p(\mathbf{y}_i | \mathbf{p}_i), \tag{2.4}$$

where

$$\begin{aligned} p(\mathbf{y}_i | \mathbf{p}_i) &= \sum_{\mathbf{z}_i : \mathbf{A}_i \mathbf{z}_i = \mathbf{y}_i} p(\mathbf{y}_i, \mathbf{z}_i | \mathbf{p}_i) \\ &= \sum_{\mathbf{z}_i : \mathbf{A}_i \mathbf{z}_i = \mathbf{y}_i} \underbrace{\binom{n_i}{z_{i1}, \ldots, z_{iH}} p_{i1}^{z_{i1}} \cdots p_{iH}^{z_{iH}}}_{p(\mathbf{z}_i | \mathbf{p}_i)} \underbrace{\mathbb{1}(\mathbf{y}_i = \mathbf{A}_i \mathbf{z}_i)}_{p(\mathbf{y}_i | \mathbf{z}_i)}. \end{aligned} \tag{2.5}$$

The summation in (2.5) requires us to enumerate, for each $i = 1, \ldots, N$, the *feasible set* of solutions

$$\mathcal{F}(\mathbf{A}_i, \mathbf{y}_i) := \{\mathbf{z} \in \mathbb{Z}_{\ge 0}^H : \mathbf{A}_i \mathbf{z} = \mathbf{y}_i\}, \tag{2.6}$$

where $\mathbb{Z}_{\ge 0}^H$ is the space of H-dimensional non-negative integer vectors. We find all elements of the feasible set with a branch-and-bound algorithm (electronic supplementary material, S1). This algorithm is run once prior to parameter inference.

2.2. Gaussian process prior specification

To account for the correlation between haplotype prevalences for different studies, we model the haplotype prevalences as a softmax transformation of H independent GPs:

$$\begin{aligned} p_{ij} &= \frac{\exp(f_j(\mathbf{x}_i))}{\exp(f_1(\mathbf{x}_i)) + \cdots + \exp(f_H(\mathbf{x}_i))} \\ &\text{for } i = 1, \ldots, N, \ j = 1, \ldots, H \end{aligned} \tag{2.7}$$

and

$$f_j(\mathbf{x}_1), \ldots, f_j(\mathbf{x}_N) \sim \mathrm{N}(m_j(\mathbf{X}), C_j(\mathbf{X}, \mathbf{X})) \quad \text{for } j = 1, \ldots, H, \tag{2.8}$$

where $\mathbf{p}_i$ are the haplotype prevalences of population i, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ are the spatio-temporal coordinates and covariates for each study, and f_j is the jth GP whose mean function and covariance function are m_j and C_j respectively. The vector $m_j(\mathbf{X})$ denotes the concatenation of $(m_j(\mathbf{x}_1), \ldots, m_j(\mathbf{x}_N))$, and $C_j(\mathbf{X}, \mathbf{X})$ is a matrix whose (i, i')th entry is $C_j(\mathbf{x}_i, \mathbf{x}_{i'})$. The mean and covariance functions are further parametrized by GP hyperparameters $\boldsymbol{\theta}$ (figure 1).

For each data point $i = 1, \ldots, N$, our covariates $\mathbf{x}_i = (\lambda_i, \phi_i, t_i, r_i)$ consist of the longitude λ_i, latitude ϕ_i, median year of study t_i, and *P. falciparum* parasite rate r_i (as estimated from the Malaria Atlas Project [13]). We assume that the mean function varies linearly with the parasite rate:

$$m_j(\mathbf{x}_i) = \mu_j + \beta_j r_i, \tag{2.9}$$

where μ_j is a baseline mean value and β_j quantifies the effect of parasite rate on the prevalence of haplotype j. We choose our covariance function from the Gneiting class of covariance functions on a sphere [14], along with a white noise term:

$$C_j(\mathbf{x}_i, \mathbf{x}_{i'}) = s_j^2 \left(1 + \frac{(t_i - t_{i'})^2}{\tau_j^2} + \frac{d_{\mathrm{GC}}(\mathbf{x}_i, \mathbf{x}_{i'})}{\delta_j}\right)^{-1} + \sigma^2 \mathbb{1}(i = i'), \tag{2.10}$$

where s_j^2 is the spatio-temporal variance, τ_j is the timescale parameter, δ_j is the lengthscale parameter, σ^2 is the noise variance, $d_{\mathrm{GC}}(\mathbf{x}_i, \mathbf{x}_{i'})$ is the great circle distance (in degrees) between data points i and i', and $\mathbb{1}(\cdot)$ is the indicator function. Other Gneiting covariance functions are possible, but we choose a simple one that resembles the rational quadratic covariance function. Note that all hyperparameters except for σ^2 are haplotype-specific. We place weakly informative priors on the GP hyperparameters $\boldsymbol{\theta} = \{\mu_j, \beta_j, s_j, \tau_j, \delta_j\}_{j=1}^H \cup \{\sigma\}$; see electronic supplementary material, S2, for details.

2.3. Bayesian inference

We perform Bayesian inference using Markov chain Monte Carlo (MCMC) to obtain the posterior distribution

$$\begin{aligned} &p(\mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta} | \mathbf{Y}) \\ &\quad \propto p(\mathbf{Y} | \mathbf{p}_1, \ldots, \mathbf{p}_N) p(\mathbf{p}_1, \ldots, \mathbf{p}_N | \boldsymbol{\theta}) p(\boldsymbol{\theta}), \end{aligned} \tag{2.11}$$

where $\mathbf{Y}$ denotes the collection of all observed data $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, the likelihood $p(\mathbf{Y} \mid \mathbf{p}_1, \ldots, \mathbf{p}_N)$ is defined in (2.4) and (2.5), the prior $p(\mathbf{p}_1, \ldots, \mathbf{p}_N \mid \boldsymbol{\theta})$ in §2.2 and the hyperprior $p(\boldsymbol{\theta})$ in electronic supplementary material, S2. We use the No-U-Turn sampler (NUTS) [15] to perform MCMC sampling, which uses gradient information of the logarithm of (2.11) to produce posterior chains of low autocorrelation. We run 5 MCMC chains each with 1000 iterations, discarding the first half as burn-in iterations. After MCMC sampling, we produce predictive maps of haplotype prevalences on grid cells of size $0.2^\circ \times 0.2^\circ$ across sub-Saharan Africa for each year between 2000 and 2020. Let $\mathbf{p}^*$ denote the vector of haplotype prevalences for an arbitrary grid cell with covariates $\mathbf{x}^*$. To obtain samples from the posterior $p(\mathbf{p}^* \mid \mathbf{Y})$, we draw samples from the distribution $p(\mathbf{p}^* \mid \mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta})$, where the values of $\mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta}$ are taken from the posterior samples output by NUTS. This is justified by the fact that

$$p(\mathbf{p}^* | \mathbf{Y}) = \int p(\mathbf{p}^* | \mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta}) p(\mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta} | \mathbf{Y}) \, \mathrm{d}\mathbf{p}_1, \cdots, \mathrm{d}\mathbf{p}_N \, \mathrm{d}\boldsymbol{\theta}. \tag{2.12}$$

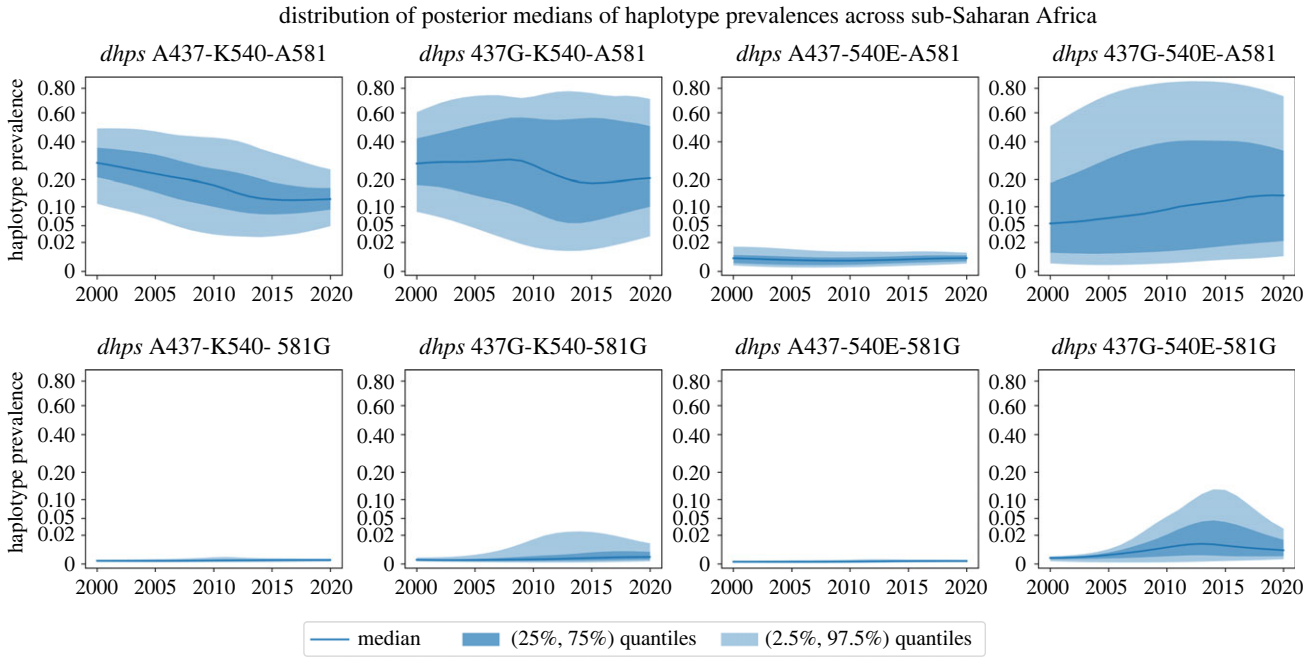

Figure 2. Distribution (over grid cells of study area) of the posterior medians of $H = 8$ prevalences of *dhps* haplotypes from year 2000 to year 2020. The dark blue line represents the median over the grid cells of our study area, the dark blue band represents the 50% interval and the light blue band the 95% interval. Posterior medians in all panels are presented on a square root scale.

Based on GP theory [16], the distribution $p(\mathbf{p}^* \mid \mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta})$ follows a normal distribution with the softmax transformation (2.7) applied, allowing the posterior predictive distribution $p(\mathbf{p}^* \mid \mathbf{Y})$ to be exactly sampled given samples of the posterior distribution $p(\mathbf{p}_1, \ldots, \mathbf{p}_N, \boldsymbol{\theta} \mid \mathbf{Y})$.

3. Results

We fit our latent multinomial GP model to SP resistance data in Africa, where it is common for parasites to have multiple SP-resistant mutations on the *dhps* gene and studies may report the number of samples with mutations differently. Here, we consider $G = 3$ mutations on the *dhps* gene, namely A437G, K540E and A581G, leading to $H = 8$ distinct full haplotypes. We use a dataset collated by the Worldwide Antimalarial Resistance Network [12], as detailed in electronic supplementary material, S3. This dataset reports a total of six realized haplotypes, listed in (2.2), although each data point typically does not report all realized haplotypes. Our primary goal is to obtain Bayesian estimates of the prevalences (i.e. multinomial probabilities) of the 8 full haplotypes across sub-Saharan Africa over the duration of interest 2000–2020.

The total computational time for preprocessing and MCMC was 7.9 h. We achieve a potential scale reduction factor [17] of $\hat{R} < 1.02$ for all parameters, and each haplotype prevalence had an effective sample size greater than 500. The MCMC chains exhibit good mixing, as indicated by the representative trace plots shown in electronic supplementary material, figure S1.

We divide the region of interest into a $0.2^\circ \times 0.2^\circ$ grid, where we define our area of interest to be the region where the Malaria Atlas Project maps malaria transmission [13,18]. For each spatio-temporal coordinate $\mathbf{x}^*$ of a grid cell and year, we find the posterior predictive distribution of the full haplotype prevalences at $\mathbf{x}^*$. Figure 2 shows the distribution of posterior median prevalences over the region of interest for all eight haplotypes during 2000–2020. Over this region, of the $H = 8$ haplotypes, three have very low prevalence over the entire duration of interest: A437-540E-A581, A437-K540-581G and A437-540E-581G.

We also provide visual summaries of the spatial distribution of prevalences of 5 selected haplotypes (the three low prevalence haplotypes from figure 2 are excluded) and their change over time. Figure 3 shows the median and standard deviation summaries of the posterior predictive distributions over the region of interest in the years 2000, 2010 and 2020. The results presented in this figure are broadly consistent with literature in that the vast majority of mutant *dhps* haplotypes have the A437G mutation [19]. The spatial patterns shown by our heatmaps agree with the results of Naidoo & Roper [20], who report the double mutation A437G-K540E as the most prevalent mutant haplotype in East Africa (third column), and also the single mutation A437G as the most prevalent mutant haplotype in West and Central Africa (second column).

In the case where the prevalence of an individual mutation (A437G, K540E or A581G) is of interest, these can still be captured as outputs of our model by simply summing over the halplotypes that include the mutation. In figure 4, the spatial distributions of A437G, K540E and A581G are summarized with posterior median (top row) and standard deviation (bottom row) in 2020. These results are broadly consistent with results shown in Flegg *et al.* [6] for the same three markers in 2020.

To assess the predictive utility of our model, we rerun the inference 10 times, each time with a different 10% of the dataset withheld from inference. For each data point of the full dataset, the haplotype prevalences are predicted by the posterior median obtained from the inference instance that did not include the data point. Details of this model validation procedure are given in electronic supplementary material, S4. We report in table 2 the mean error (measure of bias) and mean absolute error (measure of average discrepancy)

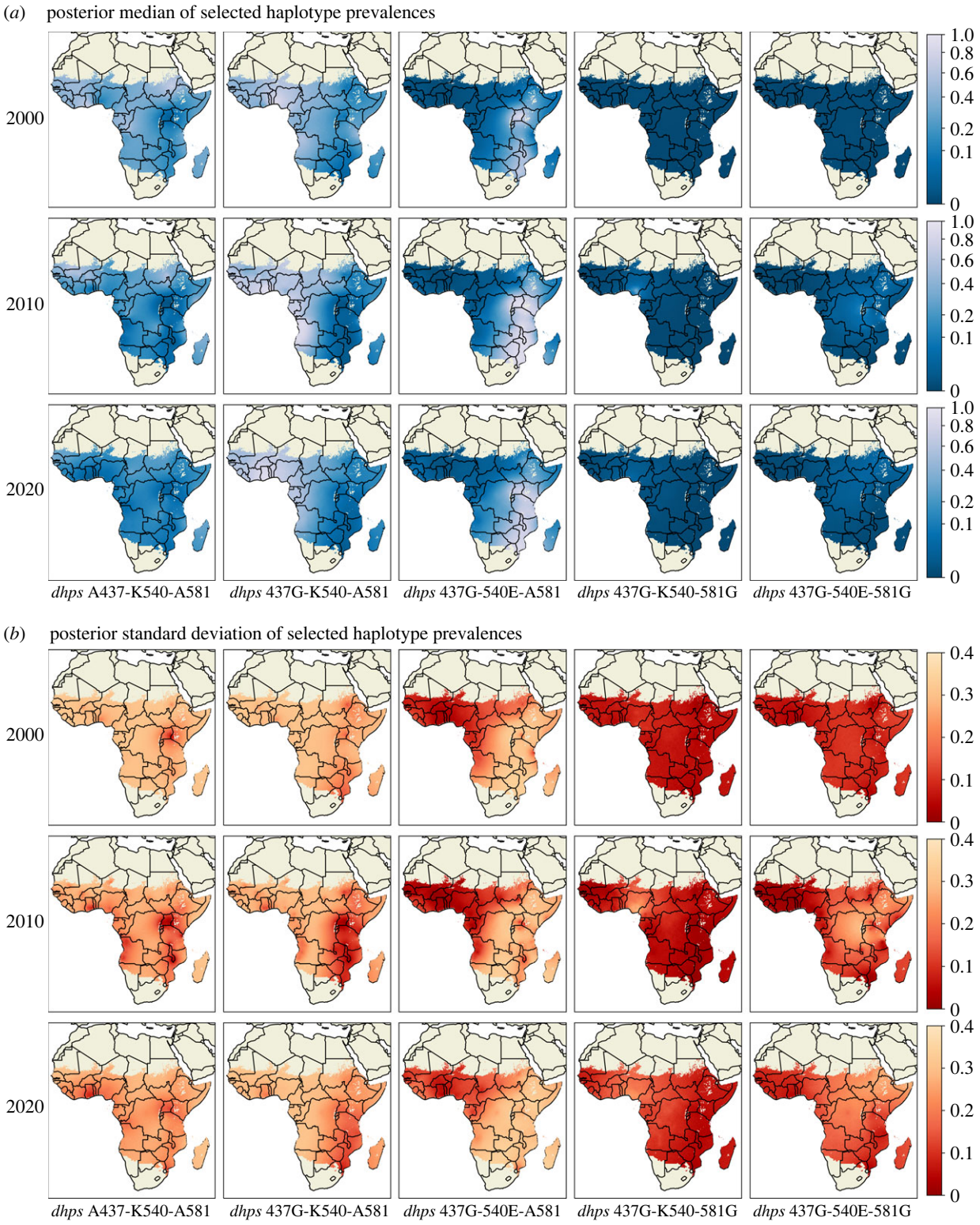

Figure 3. (*a*) Posterior median and (*b*) posterior standard deviation for prevalences of selected *dhps* haplotypes in years 2000 (top row), 2010 (middle row) and 2020 (bottom row).

between predictive median and observed prevalences for each realized haplotype. Means for each realized haplotype are taken over the data points that report the count of that realized haplotype. For the realized haplotypes that involve one mutation only, we compare our errors to those obtained by Flegg *et al.* [6], who performed spatio-temporal mapping in the same study area for individual *dhps* mutations separately. Flegg *et al.* used the same 10-fold cross-validation approach as us; see table 3 in [6] for their results. Our mean absolute errors are comparable to those of Flegg *et al.* and the direction of bias (i.e. sign of mean error) concurs for all three mutations. However, the mean errors we report have larger magnitudes, possibly due to our dataset being smaller than the dataset used in [6]. Nevertheless, there is good agreement between the predictive median and observed prevalences; see electronic supplementary material, figure S2, for scatterplots comparing the predictive medians to the corresponding observed prevalences.

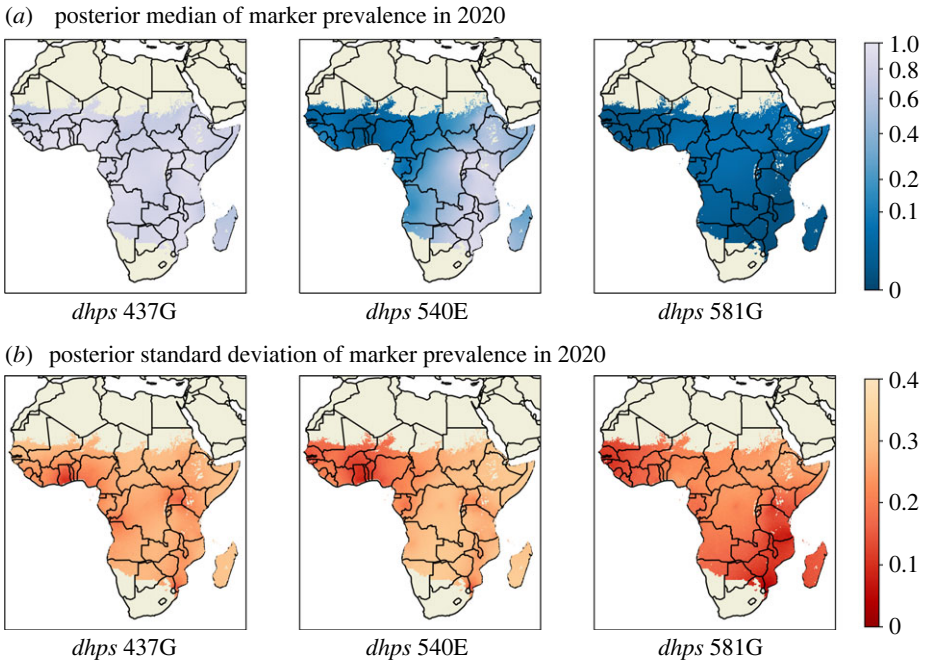

Figure 4. Posterior summary for prevalences of selected *dhps* individual markers in year 2020. The top row shows the posterior median for *dhps* 437, 540, 581; the bottom row shows the posterior standard deviation for *dhps* 437, 540, 581.

Table 2. Number of data points that report the counts for each realized haplotype, and the corresponding mean error and mean absolute error between predictive median and observed prevalences.

realized haplotype	no. data points	mean error	mean absolute error
*dhps*437G	208	0.0389	0.1158
*dhps*540E	214	0.0096	0.0731
*dhps*581G	154	−0.0302	0.0553
*dhps*437G-540E	65	−0.0060	0.0451
*dhps*437G-540E-581G	35	−0.0518	0.0778
*dhps*437G-540E-A581	33	−0.0150	0.0758

4. Discussion

In this paper, we develop a spatio-temporal geostatistical model to infer for the first time the prevalence of multi-marker drug-resistant malaria. We illustrate the utility of this new model for SP, which is a commonly used drug for intermittent preventive treatment of malaria in pregnancy, children and infants. Since drug-resistant haplotypes and markers are often used as a proxy for treatment efficacy, these maps can help inform antimalarial drug policies. Our methods take on a Bayesian approach, which are able to quantify uncertainties about the prevalence of the drug resistance haplotypes. A benefit of quantifying uncertainty is its use in optimizing sampling strategies for future monitoring of drug resistance [21].

Existing geostatistical methods in spatio-temporal modelling of antimalarial drug resistance use a binomial likelihood [3,6,21,22], which can only infer the prevalence of a single category, e.g. prevalence of one mutation, or prevalence of the wild-type haplotype. Our models are capable of handling multiple haplotypes simultaneously by using a multinomial likelihood, leading to more refined inference about drug resistance. This has not been done in previous work, as not all studies provide data on full haplotypes; studies may only report on a subset of mutations, or group haplotypes by the number of mutations present, or provide counts for each mutation separately. We were able to handle these types of partially reported data by using a latent multinomial model that treats each reported category as a subset of all full haplotypes. Although the counts of each full haplotype are not all experimentally determined, our approach of enumerating all possible latent counts allows us to leverage the partially reported data for inferring the prevalences of the full haplotypes.

One limitation of our work presented here is that sampling bias may be present due to population heterogeneity. For example, since SP is commonly used for intermittent preventive treatment of malaria, many studies of SP resistance take blood samples from pregnant people. Bias may also arise from the choice of mutations reported. If the prevalence of a mutation is very low, it is less likely to be reported by a study. This may lead to an overestimation in the prevalence of mutations that are not often reported. Another limitation is a lack of model checking to verify whether our model fits the data adequately. Since the haplotype categories reported are inconsistent across studies, it is not straightforward as to what model checking procedures should be applied.

A possible extension is to include more covariates for the GP model. Of biological interest are covariates related to drug pressure, such as treatment-seeking rates [23]. However, using more covariates implies that more model parameters need to be inferred. Our current MCMC approach is already computationally expensive, an issue that may be exacerbated by the inclusion of more covariates. For a dataset with more markers and/or larger pools, our enumeration approach may become infeasible, as there are too many possible latent count solutions to enumerate. In this case, we can instead treat the latent counts as model parameters to be sampled during

MCMC using a custom MCMC sampler [10] based on Markov bases [24]. MCMC sampling of the latent counts cannot be performed by gradient-based samplers such as NUTS, as they cannot handle discrete model parameters. This is particularly relevant if the *dhfr* gene is to be included in future analyses. The computational feasibility of such analyses depends on the number of reported haplotypes, and on the number of full haplotypes used in the statistical model. We illustrate these ideas in electronic supplementary material, S5, through a case study based on molecular data collected from India [25], focusing on a *dhfr* + *dhps* quintuple mutation that is associated with clinical failure of SP [26]. There is also more work to do in extending the model to consider dependent GPs, as it is possible that the prevalences of different haplotypes are related to each other, using for example the linear model of coregionalization [27] or the semiparametric latent factor model [28].

At present, the World Health Organization provides recommendations for implementing intermittent preventive treatment in pregnancy with SP based on the prevalence of the *dhps* K540E and A581G mutations [29]. Although it is known that different *dhps* haplotypes confer different degrees of SP resistance [20], these recommendations are based on the prevalence of individual mutations, rather than that of full haplotypes. One possible reason is that there are no existing methods in the literature to infer the prevalence of full haplotypes from partially reported data. We address this gap in the literature by describing how a latent multinomial GP model can be used to produce spatio-temporal maps of these prevalences. The results we present in this paper are able to quantify the spread of various drug-resistant haplotypes, and provide uncertainty estimates that can help optimize sampling strategies for future monitoring of antimalarial drug resistance.

Data accessibility. The dataset supporting this article is available from the Wwarn repository: www.wwarn.org/tracking-resistance/sp-molecular-surveyor [30]. The code used for analysis is available from the Zenodo repository: https://zenodo.org/records/10354808 [31].

Further details of our methods and results are provided in electronic supplementary material [32].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. Y.S.F.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; J.A.F.: conceptualization, funding acquisition, investigation, methodology, supervision, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. J.A.F.'s research is supported by the ARC (grants no. DP200100747 and FT210100034) and the National Health and Medical Research Council (grant no. APP2019093).

Acknowledgements. The authors thank three anonymous reviewers for their helpful comments.

References

1. Sibley CH *et al.* 2001 Pyrimethamine–sulfadoxine resistance in *Plasmodium falciparum*: what next? *Trends Parasitol.* **17**, 582–588. (doi:10.1016/S1471-4922(01)02085-2)
2. Nsanzabana C, Djalle D, Guérin PJ, Ménard D, González IJ. 2018 Tools for surveillance of antimalarial drug resistance: an assessment of the current landscape. *Malar. J.* **17**, 75. (doi:10.1186/s12936-018-2185-9)
3. Flegg JA, Patil AP, Venkatesan M, Roper C, Naidoo I, Hay SI, Sibley CH, Guerin PJ. 2013 Spatiotemporal mathematical modelling of mutations of the *dhps* gene in African *Plasmodium falciparum*. *Malar. J.* **12**, 249. (doi:10.1186/1475-2875-12-249)
4. Deutsch-Feldman M *et al.* 2019 The changing landscape of *Plasmodium falciparum* drug resistance in the Democratic Republic of Congo. *BMC Infect. Dis.* **19**, 872. (doi:10.1186/s12879-019-4523-0)
5. Amimo F, Lambert B, Magit A, Sacarlal J, Hashizume M, Shibuya K. 2020 *Plasmodium falciparum* resistance to sulfadoxine-pyrimethamine in Africa: a systematic analysis of national trends. *BMJ Global Health* **5**, e003217. (doi:10.1136/bmjgh-2020-003217)
6. Flegg JA *et al.* 2022 Spatiotemporal spread of *Plasmodium falciparum* mutations for resistance to sulfadoxine-pyrimethamine across Africa, 1990–2020. *PLoS Comput. Biol.* **18**, e1010317. (doi:10.1371/journal.pcbi.1010317)
7. Gesase S *et al.* 2009 High resistance of *Plasmodium falciparum* to sulphadoxine/pyrimethamine in northern Tanzania and the emergence of dhps resistance mutation at codon 581. *PLoS ONE* **4**, e4569. (doi:10.1371/journal.pone.0004569)
8. Harrington W, Mutabingwa T, Muehlenbachs A, Sorensen B, Bolla M, Fried M, Duffy P. 2009 Competitive facilitation of drug-resistant *Plasmodium falciparum* malaria parasites in pregnant women who receive preventive treatment. *Proc. Natl Acad. Sci. USA* **106**, 9027–9032. (doi:10.1073/pnas.0901415106)
9. Link WA, Yoshizaki J, Bailey LL, Pollock KH. 2010 Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics* **66**, 178–185. (doi:10.1111/j.1541-0420.2009.01244.x)
10. Foo YS, Flegg JA. 2023 Haplotype frequency inference from pooled genetic data with a latent multinomial model. (https://arxiv.org/abs/2308.16465)
11. Triglia T, Wang P, Sims PF, Hyde JE, Cowman AF. 1998 Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *EMBO J.* **17**, 3807–3815. (doi:10.1093/emboj/17.14.3807)
12. Infectious Diseases Data Observatory. 2022 SP Molecular Surveyor. See https://www.wwarn.org/tracking-resistance/sp-molecular-surveyor (accessed 18 June 2023).
13. Malaria Atlas Project. 2023 Data. See https://malariaatlas.org/ (accessed 20 May 2023).
14. Porcu E, Furrer R, Nychka D. 2021 30 years of space–time covariance functions. *WIREs Comput. Stat.* **13**, e1512. (doi:10.1002/wics.1512)
15. Hoffman MD, Gelman A. 2014 The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623.
16. Rasmussen CE, Williams CKI. 2005 *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. New York, NY: The MIT Press.
17. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. 2021 Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16**, 667–718. (doi:10.1214/20-BA122110.1214/20-BA1221)
18. Weiss DJ *et al.* 2019 Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *The Lancet* **394**, 322–331. (doi:10.1016/S0140-6736(19)31097-9)
19. Vinayak S *et al.* 2010 Origin and evolution of sulfadoxine resistant *Plasmodium falciparum*. *PLoS Pathog.* **6**, e1000830. (doi:10.1371/journal.ppat.1000830)
20. Naidoo I, Roper C. 2013 Mapping 'partially resistant', 'fully resistant', and 'super resistant' malaria. *Trends Parasitol.* **29**, 505–515. (doi:10.1016/j.pt.2013.08.002)
21. Grist EPM *et al.* 2016 Optimal health and disease management using spatial uncertainty: a geographic characterization of emergent

artemisinin-resistant *Plasmodium falciparum* distributions in Southeast Asia. *Int. J. Health Geogr.* **15**, 37. (doi:10.1186/s12942-016-0064-6)
22. Tun KM *et al.* 2015 Spread of artemisinin-resistant *Plasmodium falciparum* in Myanmar: a cross-sectional survey of the K13 molecular marker. *Lancet Infect. Dis.* **15**, 415–421. (doi:10.1016/S1473-3099(15)70032-0)
23. Nguyen M *et al.* 2023 Trends in treatment-seeking for fever in children under five years old in 151 countries from 1990 to 2020. *PLoS Global Public Health* **3**, e0002134. (doi:10.1371/journal.pgph.0002134)
24. Schofield MR, Bonner SJ. 2015 Connecting the latent multinomial. *Biometrics* **71**, 1070–1080. (doi:10.1111/biom.12333)
25. Ahmed A, Bararia D, Vinayak S, Yameen M, Biswas S, Dev V, Kumar A, Ansari MA, Sharma YD. 2004 *Plasmodium falciparum* isolates in India exhibit a progressive increase in mutations associated with sulfadoxine-pyrimethamine resistance. *Antimicrob. Agents Chemother.* **48**, 879–889. (doi:10.1128/AAC.48.3.879-889.2004)
26. Kublin JG *et al.* 2002 Molecular markers for failure of sulfadoxine–pyrimethamine and chlorproguanil–dapsone treatment of *Plasmodium falciparum* malaria. *J. Infect. Dis.* **185**, 380–388. (doi:10.1086/338566)
27. Journel AG, Huijbregts CJ. 1978 *Mining geostatistics*. London, UK: Academic Press.
28. Teh YW, Seeger M, Jordan MI. 2005 Semiparametric latent factor models. *Proc. Mach. Learn. Res.* **R5**, 333–340.
29. World Health Organization. 2013 *Meeting report of the Evidence Review Group on intermittent preventive treatment (IPT) of malaria in pregnancy*. Geneva, Switzerland: WHO.
30. Foo YS, Flegg JA. 2024 A spatio-temporal model of multi-marker antimalarial resistance. Wwarn repository. (https://www.wwarn.org/tracking-resistance/sp-molecular-surveyor)
31. Foo YS, Flegg JA. 2024 A spatio-temporal model of multi-marker antimalarial resistance. Zenodo. (https://zenodo.org/records/10354808)
32. Foo YS, Flegg JA. 2024 A spatio-temporal model of multi-marker antimalarial resistance. Figshare. (doi:10.6084/m9.figshare.c.6992018)