

RESEARCH

Open Access



# Bayesian spatio-temporal distributed lag modeling for delayed climatic effects on sparse malaria incidence data

Chawarat Rotejanaprasert<sup>1,2\*</sup>, Nattwut Ekampirat<sup>2</sup>, Prayuth Sudathip<sup>3</sup> and Richard J. Maude<sup>2,4,5,6</sup>

## Abstract

**Background:** In many areas of the Greater Mekong Subregion (GMS), malaria endemic regions have shrunk to patches of predominantly low-transmission. With a regional goal of elimination by 2030, it is important to use appropriate methods to analyze and predict trends in incidence in these remaining transmission foci to inform planning efforts. Climatic variables have been associated with malaria incidence to varying degrees across the globe but the relationship is less clear in the GMS and standard methodologies may not be appropriate to account for the lag between climate and incidence and for locations with low numbers of cases.

**Methods:** In this study, a methodology was developed to estimate the spatio-temporal lag effect of climatic factors on malaria incidence in Thailand within a Bayesian framework. A simulation was conducted based on ground truth of lagged effect curves representing the delayed relation with sparse malaria cases as seen in our study population. A case study to estimate the delayed effect of environmental variables was used with malaria incidence at a fine geographic scale of sub-districts in a western province of Thailand.

**Results:** From the simulation study, the model assumptions which accommodated both delayed effects and excessive zeros appeared to have the best overall performance across evaluation metrics and scenarios. The case study demonstrated lagged climatic effect estimation of the proposed modeling with real data. The models appeared to be useful to estimate the shape of association with malaria incidence.

**Conclusions:** A new method to estimate the spatiotemporal effect of climate on malaria trends in low transmission settings is presented. The developed methodology has potential to improve understanding and estimation of past and future trends in malaria incidence. With further development, this could assist policy makers with decisions on how to more effectively distribute resources and plan strategies for malaria elimination.

**Keywords:** Spatiotemporal, Malaria, Bayesian, Lag effect, Weather

## Background

Among the parasitic diseases worldwide, malaria is one of the most prevalent [1]. It is caused by protozoan parasites of the genus *Plasmodium*. In the Greater Mekong Subregion (GMS), the predominant species are *P. falciparum*

and *P. vivax*. As the species most likely to cause severe complications and death, and lacking a dormant liver stage, *P. falciparum* has been the main focus of elimination plans among countries in the region. Migration and agricultural activities along the country's borders have been studied and found to be associated with malaria risk. Either side of the Thailand-Myanmar border has consistently been an area with relatively high transmission for many years [2]. With intense efforts to eliminate

\*Correspondence: chawarat.rot@mahidol.ac.th

<sup>2</sup> Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand

Full list of author information is available at the end of the article



malaria by 2030, the number of *P. falciparum* incidence has markedly decreased in recent years across the GMS [3]. This includes on the Thai side [4] and in Kayin State in Myanmar [5]. In many areas of the Greater Mekong Sub-region, malaria is now narrowed down to small foci of transmission and there has been an intensification of efforts by National Malaria Control Programs and partners to better understand and stop residual transmission in order to achieve elimination.

The malaria parasite is transmitted from human to human via the bite of infected female mosquitoes of the genus *Anopheles*. The malaria case distribution and dynamics have been found to be closely related to environmental factors, particularly in high transmission areas in Sub-Saharan Africa [6]. Studies have shown that both mosquito species and malaria parasites are very sensitive to weather conditions [6–11]. For example, moderate rainfall can create mosquito breeding sites while temperature affects the rate of development of the mosquito larvae and influences mosquito survival rates [12, 13]. In addition, higher temperatures also accelerate multiplication of the *Plasmodium* parasites inside the vectors [14]. Suitable climatic conditions can create circumstances appropriate for malaria transmission in endemic areas. Therefore, how malaria incidences change as a result of climatic variability is an important understanding for effective malaria control planning and activities [15]. The relationship between climate and malaria incidence is complex with inconsistent findings perhaps due to regional variations and limited availability of suitable methods [16].

To investigate the relationship between meteorological variables and malaria incidence, it is sensible to account for biology of the transmission process. It has been shown that the relationship between climate and malaria is stronger for *P. falciparum* than for *P. vivax* [17]. This is likely to be due to a proportion of *P. vivax* cases being recurrences of dormant liver stage parasites (not found in falciparum infections) rather than being transmitted by mosquitoes. Epidemiologically there is a lag between changes in environmental factors and malaria transmission due to both the life cycle of mosquitoes (only mature adult females transmit malaria) and incubation of parasites in the mosquitoes (which are transmitted once they develop into sporozoites). This forms a lagged time-varying distribution of the association. Therefore, the lag effect should then be addressed in analyses. To account for the lagged correlation, distributed lag nonlinear modeling (DLNM) has been shown to be a valuable and effective tool [18–20].

Distributed lag models (DLMs) have been widely applied in epidemiological studies to estimate associations between environmental variables and health

outcome of interest at different lagged periods (for example please see [20–23]). Although the use of DLMs to quantify the lag effect of climatic factors and malaria incidences has been used (e.g., [10, 11, 16, 24]), currently few studies have investigated the possibility of spatio-temporal variation in health outcomes within the lagged regression modeling. Contributors to such spatial variation can include exposure characteristics which could be spatially varying environmental composition and spatial differences in exposure measurements [25]. To account for spatial variation, geostatistical point process methods have been used to introduce spatial correlation between lagged regression parameters corresponding to different discrete spatial regions while defining correlation as a function of distance between region centroids [26]. However, this distance may be inappropriate when the spatial regions are oddly shaped and/or their sizes vary greatly [27]. Recently a spatially varying distributed lag model at individual level has been developed with application to an air pollution and term low birth weight study [25]. However, since the number of *P. falciparum* malaria cases of countries in GMS has decreased and distribution of cases is increasingly sparse as areas progress towards elimination, the excess of zero cases is another analytical challenge that has not been considered in previous modeling.

Excessive zeros commonly occur and are often encountered in biostatistical and epidemiological research especially for nearly eliminated and neglected tropical diseases. Based on different data generation mechanisms, the large number of zeros can lead to overdispersion caused by a disagreement between the data and the assumed distribution. This can result in having more zeroes in our data than the proposed distribution could reasonably explain. The zero valued data should not be ignored and dropped from the analysis as they often provide important information regarding the process. In addition, having a large proportion of zeros could indicate important characteristics of the data or condition under study. Thus, probabilistic models that are capable of handling excessive zeros should be considered. Excessive-zero modeling has been frequently adopted in the disease mapping literature (see for example [28–31]). To account for the temporal lagged effect, an autoregressive integrated moving average (ARIMA) model has been used to model correlated residual in relative malaria risk [32]. However, that method does not directly address the collinearity in climate covariates which is a known issue [33, 34] and could be accounted for using DLMs.

Therefore, in this study we aimed to develop a framework to estimate the temporal lag effect of climatic factors on *P. falciparum* malaria incidence in Thailand. Whereas the relationship of multiple malaria species has

been investigated [35–37], in this work we focused on the association of the single species with climatic variables which is a current concern of disease control and elimination in the area. The sub-district data were from passive surveillance at level which there have not been analyses well performed at this fine spatial scale. We described research motivation in the context of malaria elimination and provided the basic model used in spatio-temporal analysis. The concept of DLMs as a foundation of our methodology to describe a lagged effect of climate on malaria incidence was introduced. Then the additional complexity of the zero-inflation issue in our data set and spatio-temporal distributed lagged modeling for zero inflated data was described. The robustness of the proposed framework was examined through a simulation study and a case study to estimate the delayed effect of environmental variables on malaria incidence in Tak Thailand along the Thailand-Myanmar border, one of the areas with the highest malaria burden in Thailand. Finally, analytic challenges and developments are further discussed.

## Methods

### Study design

In this study we first developed a methodology to investigate the lagged distribution of climatic effects on *P. falciparum* malaria cases in western Thailand. Then a simulation study was conducted to examine the performance of the models under various parameter conditions and assessed using a number of evaluation metrics. A case study with retrospective analysis design was also provided as a real data example of lagged effects on malaria cases in an elimination setting along the Thailand-Myanmar border. Weekly malaria incidence data of Tak in 2016 obtained from the Bureau of Vector Borne Disease at the Ministry of Public Health were used to demonstrate the application of the proposed model. To quantify the association with climatic factors, data on average temperature, relative humidity and total rainfall were considered in this study and collected from the Meteorological Department from five weather stations across the region during the study period.

### Modeling for spatially sparse health data

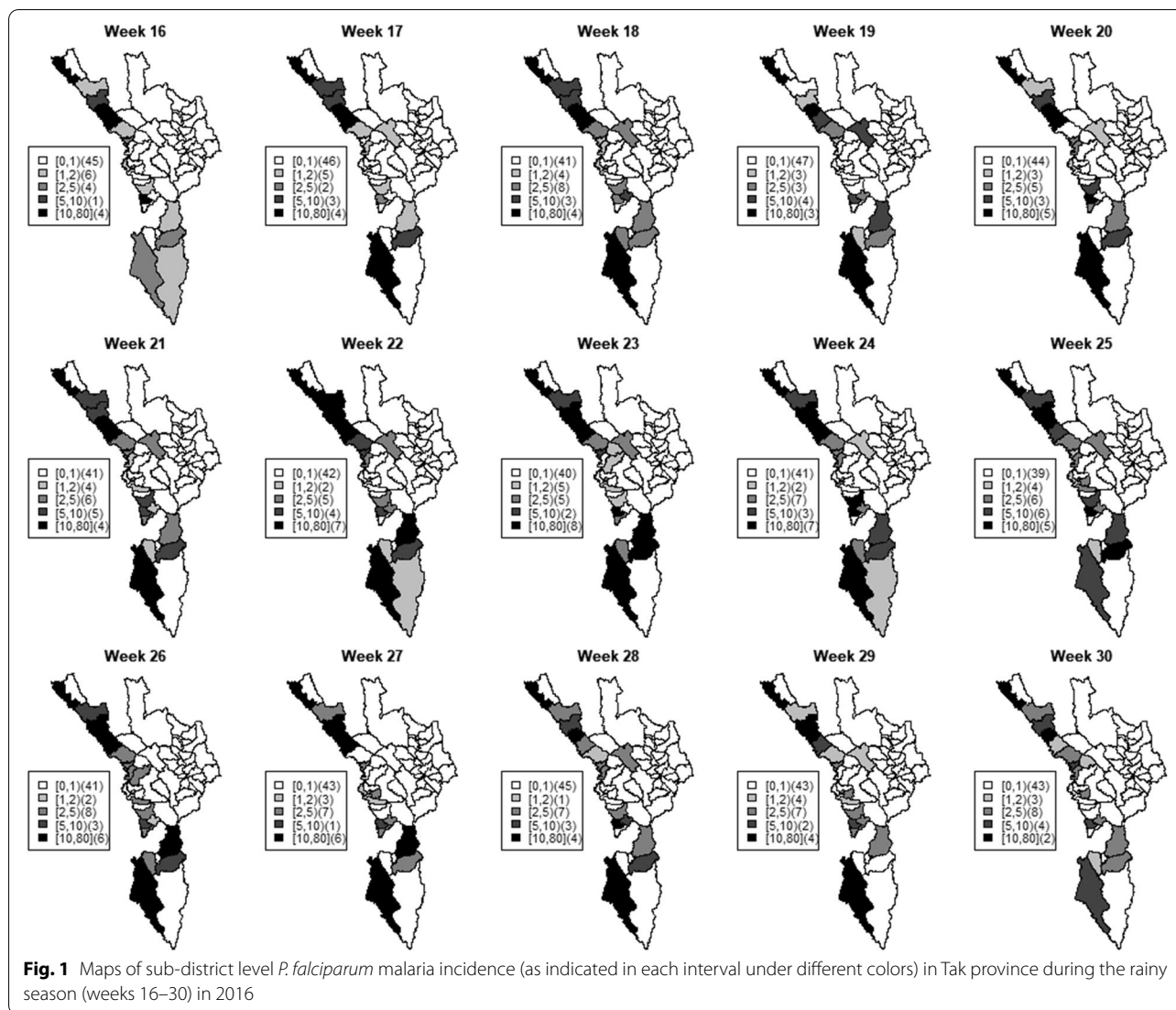
The analysis of malaria incidence over space and time has received considerable attention due to growing demand for reliable estimation in order to plan effective control activities. Modeling incidence from surveillance health data indexed at a fine spatial resolution poses specific statistical problems due to the sparse nature of the data, especially in elimination settings. Figure 1 presents the number of *P. falciparum* malaria cases at sub-district level (as indicated in each interval under different colors)

in Tak province in 2016 during weeks 16–30 which was the rainy season, the period of high malaria transmission in the region due to weather suitability. Figure 2 shows histograms of malaria incidence in the same study population. Both figures present the sparseness in our data set. In general, estimates of small areas can be sensitive to sampling variation due to relative population size to rare disease incidence among areas [38, 39]. Model-based methods have been developed to address the variability issue when mapping health outcomes at small geographical resolutions. These methods help to decrease the variability in small area estimates in sparse health data using spatial smoothing, which allows study units to integrate strength from surrounding regions to yield a more stable estimate in each area (see examples [39–41]).

The small area modeling has been widely applied on disease mapping with an assumption of Poisson likelihood [39, 42]. However, ordinary count distributions, such as regular Poisson distribution cannot adequately capture the variation due to the sparse nature of the malaria cases in our study population which also produce excessive zero cases in the data set as depicted in Figs. 1 and 2. This situation occurs primarily in disease elimination areas in which there is an abundance of zeros coupled with a heterogeneous distribution of positive counts, as these features impose competing influences on the model and potentially become problematic. In the Poisson case, for example, the high proportion of zeros tends to lower the mean parameter. While alternative count distributions, such as the negative binomial, can add flexibility by incorporating a separate heterogeneity parameter, these distributions do not always ensure adequate fit [43] and neglect of excess zeroes will bias the estimation of parameters [44].

### Spatio-temporal modeling for excessive zero malaria cases

The excessive amount of zero incidence is an analytical challenge of eliminating disease modeling. One possible explanation for those zero cases is that they occur in spatial unit where malaria transmission is not suitable to occur during the study period. This can happen for various epidemiological reasons such as vector habitat or environmental unsuitability or effective interventions. To address excessive zeros in our data, there is a broad class of spatio-temporal models to accommodate both zero inflation and nonzero counts. The two widely used approaches are hurdle and zero-inflation models. One limitation of standard models for count data is that the two types of data, zeros and the positive integers are considered to be from the same data-generating process. However, for *hurdle models*, these two mechanisms are not assumed to be the same. The basic idea is that a probability distribution governs the binary index of whether a



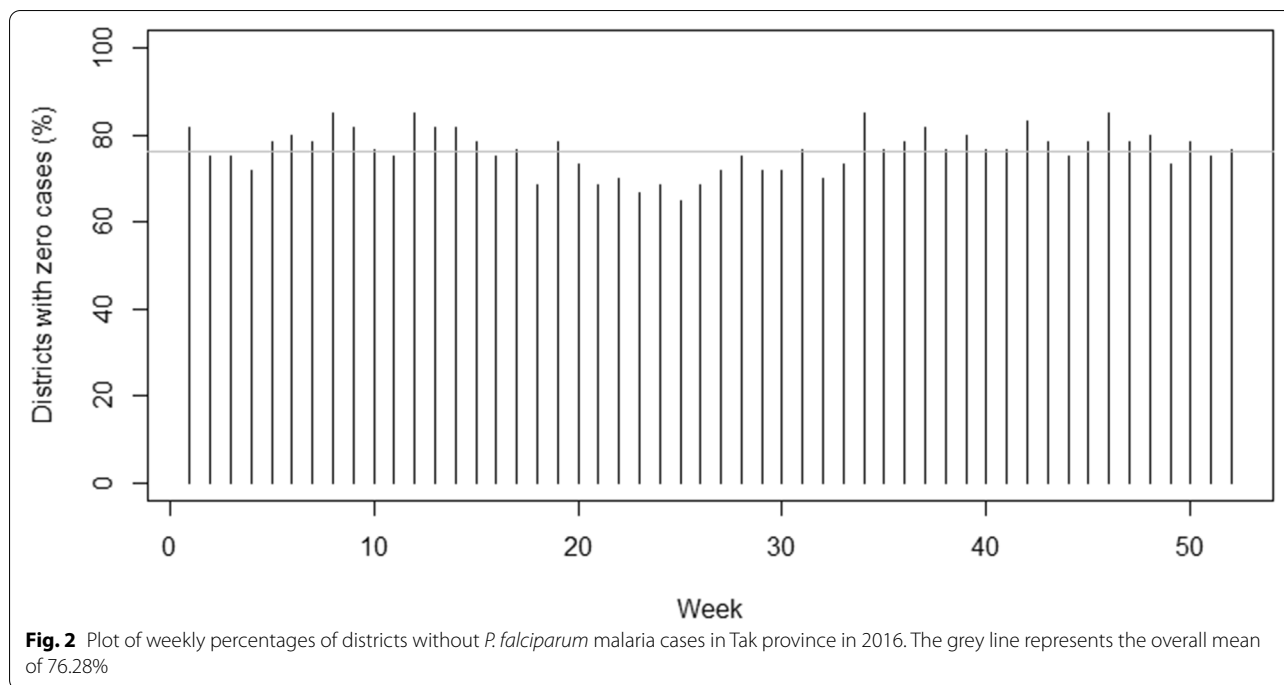
count outcome has a zero or positive integer. If the realization is positive, the hurdle model is crossed, and the conditional distribution of the positive counts is formed by a truncated-at-zero count data model. On the other hand, with zero-inflated models, the response variable is modelled as a mixture of a Bernoulli distribution (otherwise called a point mass at zero) and a base count distribution. However, for our malaria incidence data, it is perhaps more sensible to assume that our observed zero counts were from both mechanisms which could also account for asymptomatic cases. Hence, we would consider zero-inflation rather than hurdle modeling in this study.

Zero-inflation models can be seen as mixture modeling with a point mass at zero with a base distribution for the positive outcomes [30, 45]. Letting  $Y_{it}$  denote a random variable representing the malaria case count in

sub-district  $i$  in week  $t$ , the generic structure of the zero-inflation model is given by

$$Y_{it} = y_{it} \sim (1 - \pi)I_{(y_{it}=0)} + \pi f(y_{it}|\mu_{it}, \theta)I_{(y_{it}>0)}$$

where  $I_{(\cdot)}$  denotes as the indicator function;  $\pi = p(Y_{it} > 0)$  is the probability of a non-zero malaria case;  $f(y_{it}|\mu_{it}, \theta)$  is the base probability distribution with mean  $\mu_{it}$  conditioned on all parameters  $\theta$  in the model. The spatio-temporal zero-inflation model specified above partitions the zeros into two types: structural or inaccessibility zeros. Hence, the true (inaccessibility) zeros part can, for instance, be represented as those that occur because they were not susceptible to malaria infection (e.g. residence in areas unsuitable for transmission). The structural part is to generate chance or sampling zeros which form those that occur by chance among those with susceptibility. So



$\pi$  can be expressed as the “at-risk” probability which can also be interpreted as the probability of belonging to an at-risk sub-population [46, 47].

**Lagged effects of potential climate factors on malaria incidence**

The association between malaria incidence and different environmental factors is complex with inconsistent findings between studies. Besides the variations due to excessive zeros which can pose an analytical challenge which was addressed in the previous section, the inconsistent evidence of the association may also result from invalid statistical assumptions including from the misspecification of the single fixed lag. The comprehensive lag pattern for climatic factors which has been thoroughly examined previously should be investigated. Thus, lagged modeling may have the potential to better understand the relationship and improve forecasting of changes in malaria incidence which would help public health authorities on how to more effectively plan and distribute resources for malaria elimination.

Modeling the relationship with environmental exposures then requires the additional lag dimension of an exposure–incidence relationship, describing the time series of the effect. Thus, we further extended the spatiotemporal zero-inflated modeling in the previous section to account for the delayed effects. Distributed lag modeling has been developed and applied to many areas

(please see examples in [18, 19, 21, 22, 48]). A general specification to describe the lagged structure can be expressed as

$$h(\mu_{it}) = \beta_0 + \sum_{l=1}^L g(x_{it-l}; \beta_l) + \delta_{it}$$

where  $L$  is the maximum lag,  $\mu_{it}$  is the mean malaria incidence of each sub-district  $i$  and week  $t$ , from an assumed base distribution, i.e.  $E(Y_{it}) = \mu_{it}$ .  $h(\cdot)$  is the link function of mean incidence and  $g(\cdot)$  represents a function of the association between malaria incidence and exposures,  $x_{it}$ , with coefficient parameter vector  $\beta_l$ . The variable  $\delta_{it}$  could include random effect terms and other risk factors with linear effects specified by the related coefficients. This specification allows the effect of a single risk factor to be modeled over a period of time, using various parametric forms to describe the distributed contributions at different time lags.

This class of models has been extensively used to assess the lagged effects of environmental factors which can be done through the choice of a basis function [12]. The related basis functions comprise a set of completely known transformations of the original exposure variable that generate a new set of variables, termed basis variables though specification of  $g(\cdot)$ . The simplest formulation is an unconstrained structure, specified in a form of the inclusion a parameter for each lag using direct linear summation [21, 49]. However, this estimation of effects at each time lag can be very unstable and in turn inflate

variance due to the collinearity between exposures at lag times [50]. Then some conditions can be imposed to improve the estimation precision of the distributed lag association. For instance, a constant association within lagged time intervals [23], or a smooth temporal relation using continuous mean functions, e.g., polynomial transformations [21, 51] or splines [22] can be assumed. This family of model specifications can be applied in which the formulation of the lags along time series is modeled with an appropriate basis function. Thus, an expression of the constrained model can be generally defined as  $g(x_{it}; \beta_l) = \sum_l \beta_l x_{it-l}$ , where the coefficient vector is also a linear combination of basis variables. Hence,  $\beta_l = \sum_m^M B_{lm} a_m$  where  $B_{ij}$  is a member in the  $(L + 1) \times M$  matrix of basis variables with the particular functional bases with degree of freedom  $M$  to the time lag  $l$  vector and  $a_m$ , corresponding coefficient vector for the basis function. There are variants of the basis function as mentioned. For instance, the matrix can be defined as a diagonal matrix for the unconstrained model, or a series of polynomials or spline mean functions of period  $l$  for distributed modeling to explain the effect as a smoothed association along lagged time periods.

**Spatio-temporal distributed lag modeling for sparse areal malaria incidence data**

To investigate the association with climatic factors, we propose a spatio-temporal distributed lag space-time model for excessive zeros in our study population. To capture the variability in the positive count data, we consider a Poisson distribution as a standard assumption for areal disease mapping [39, 42]. Although alternatives such as Negative Binomial distribution have also been used to manage dispersion problems, the Poisson likelihood with random effect terms can account for the extra variation in space-time disease mapping [52]. Therefore, the proposed model can be expressed as

$$\begin{aligned}
 p(Y_{it} = 0) &= (1 - \pi_i) + \pi_i e^{-\mu_{it}} \\
 p(Y_{it} > 0) &= \pi_i \frac{\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} > 0 \\
 i &= 1, \dots, I; t = 1, \dots, T \\
 \log(\mu_{it}) &= \log(\text{pop}_i) + \delta_{it} + \sum_k^K \sum_l^L \beta_l x_{kit-l} \\
 \beta_l &= \sum_m^M B_{lm} a_m; \delta_{it} = \beta_0 + u_i + \lambda_t
 \end{aligned}$$

The conditional mean malaria incidence  $\mu_{it}$  is linked to the linear predictor through a logarithm function as the conical link function for the exponential family with the offset set as the population on the log scale in each spatial unit. Then the conditional mean and variance of the zero-inflation Poisson (ZIP) model are  $E(Y_{it}) = \pi_i \mu_{it}$  and  $Var(Y_{it}) = \pi_i \mu_{it} [1 + (1 - \pi_i) \mu_{it}]$ .  $\beta_0$  is the common intercept  $B_{lm}$  is generated from a natural spline basis matrix with degree of freedom  $J$  to the delay  $l$  vector and  $a_m$  a vector

of corresponding parameters specific for distributed lag coefficients.

For parameter estimation, we applied a fully Bayesian approach in which prior distributions for all parameters in the model need to be assumed. In general, it is important to include full random effects in space and time dimensions with interactions. However, in this case we aim to reduce the variation in the estimation and overly saturated specification could lead to severe identifiable estimates which might result in an increase in variance instead. Therefore, we rely in our analysis on a parsimonious specification using separate spatial and temporal random terms. To incorporate the spatial correlation structure and borrowing of information across neighboring sub-districts, the spatially structured effect,  $u_i$ , is assumed to follow the intrinsic conditional autoregressive (ICAR) model proposed by Besag et al. [53]. The spatial correlation is introduced by defining  $[u_i | u_j, i \neq j, \tau_i^{-1}] \sim N(\bar{u}_i, \tau_i^{-1})$ , where  $\bar{u}_i$  is the weighted average and  $\tau_i$  is its precision parameter. They are defined as  $\bar{u}_i = \frac{\sum_j u_j w_{ij}}{\sum_j w_{ij}}$ , and  $\tau_i^{-1} = \frac{\tau_u^{-1}}{\sum_j w_{ij}}$ . The ICAR model assumes maximum spatial correlation with a binary weighted matrix  $w_{ij} = 1$  for neighboring districts and  $w_{ij} = 0$  otherwise, is taken. The temporal effect  $\lambda_t$  is modelled by a first order random walk process, where each week is influenced by variability of the previous except the first. In general, a random walk is assumed to have a prior Gaussian distribution with mean as the previous time point which can be either positive or negative [39, 54, 55]. Then the temporal trend can be expressed as  $\lambda_t \sim Normal(\lambda_{t-1}, \tau_\lambda^{-1}), t > 0; \lambda_t \sim Normal(0, \tau_\lambda^{-1}), t = 0$ . The prior distribution of all precision parameter was assumed as a non-informative distribution [39, 42, 56, 57].

**Simulation study**

To evaluate the performance of the proposed spatio-temporal models, we conducted a study based on ground-truth simulation of lagged effect curves representing the delayed relation with excessive zero malaria cases as seen in our study population. The spatial unit used in the simulation was 60 sub-districts in Tak province ( $I = 60$ ). The risk factor  $x_{it}$  was set to represent the weekly series of climatic covariate as  $x_{i1} \sim Normal(1, 0.01)$ ,  $t = 1$  and  $x_{it} = \rho x_{it-1} + \epsilon_t$ ,  $\epsilon_t \sim Normal(0, 0.01)$ ,  $t > 1$  with  $\rho = 0.8$  and  $0.3$  to represent two different situations of high and low levels of collinearity, correlation between lagged variables. Collinearity is expected to produce unstable computation and increase uncertainty in model estimates depending on the strength of correlation. Results from a simulation study suggested that even low correlation might

decrease accuracies in model parameter estimation [58]. Therefore, we also investigated the model behaviors at different levels of collinearity based on previous research and simulations [58–60]. Then we generated the coefficients,  $\beta_l$ , to represent two situations of lagged effects in the simulation study as depicted in Fig. 3. The first curve (Fig. 3 left) was the situation in which the effect increased with a peak in lag at 3 weeks and died out subsequently whereas the effect of the second assumed situation (Fig. 3 right) was an exponential decay over lag periods.

To assess different levels of zero-case percentages, since the overall mean of weekly zero-case percentages during the study period was 76.28% (~ 80%, please see Fig. 1), we set  $\pi_i = 0.8 \forall i$  in one simulation scenario to represent the excessive zeros in our data set which was relatively high. However, we also would like to consider another situation with a lower degree of zero inflation in our simulation study and therefore  $\pi_i = 0.4$  was set in the other simulation scenario. One hundred data sets were then generated and for each simulation replicate we created an outcome series  $y_{it}$  of weekly *P. falciparum* malaria incidence, with time period,  $t = 1, \dots, 20$  weeks, from a Poisson distribution coupled with relative risk  $\mu_{it}$ . Hence the simulated data were created as  $y_{it} \sim \text{ZI-Poisson}(\mu_{it})$  which follows

$$y_{it} \sim \begin{cases} \pi_i \times 1_{\{y_{it}=0\}} \\ (1 - \pi_i) \times \text{Poisson}(\mu_{it}); y_{it} \geq 0 \end{cases}$$

$$i = 1, \dots, 60; t = 1, \dots, 20$$

$$\log(\mu_{it}) = \log(\text{pop}_i) + \delta_{it} + \sum_{l=1}^{10} \beta_l x_{it-l}$$

$$\delta_{it} = \beta_0 + v_i + \lambda_t$$

$$x_{i1} \sim \text{Normal}(1, 0.01)$$

$$x_{it} = \rho x_{it-1} + \varepsilon_t, \varepsilon_t \sim \text{Normal}(0, 0.01), t > 1$$

where the priors were modeled as

$$\beta_0 \sim \text{Normal}(0, \tau_{\beta_0}^{-1}), u_i \sim \text{ICAR}(\tau_v^{-1}),$$

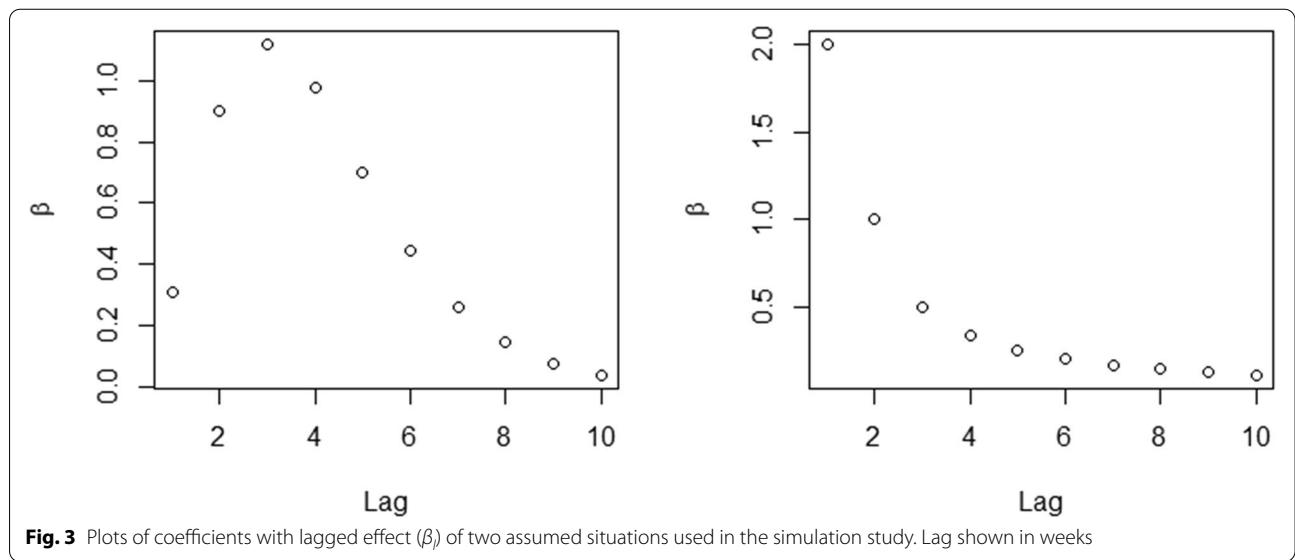
$$\lambda_t \sim \text{Normal}(\lambda_{t-1}, \tau_\lambda^{-1}), t > 0; \lambda_t \sim \text{Normal}(0, \tau_\lambda^{-1}), t = 0$$

$$\tau_{\beta_0}^{-1}, \tau_v^{-1}, \tau_\lambda^{-1} \sim \text{Unif}(0, 100)$$

To compare the performance of different model assumptions, we denoted the unconstrained model without the zero-inflation structure (one-part model) as GLM, regarded as regular generalized linear modeling with Poisson likelihood; the one-part distributed lag model was named as DLM; and zero-inflation unconstrained and zero-inflation distributed lag models were denoted ZIPGLM and ZIPDLM respectively. To fit the models, the coefficients of the unconstrained models, GLM and ZIPGLM, were fitted as  $\beta_l \sim \text{Normal}(0, 100)$  for all  $l$  ( $L = 10$ ), while DLM and ZIPDLM were fitted with three sets of degree of freedom for the basis natural spline as  $\beta_l = \sum_j B_{lj} \alpha_j$  where  $B_{lj}$  is the element  $(l, j)$  of a natural spline matrix (degrees of freedom = 3, 4, 5). The intercept was assumed to have a zero-mean Gaussian distribution whereas the prior of the probability of being at the point-mass zero was assumed to follow. The prior distribution of all precision parameter was defined as  $\tau_{\beta_0}^{-1}, \tau_v^{-1}, \tau_\lambda^{-1} \sim \text{Unif}(0, 100)$ . This allowed the maximum value of variance to be 100 which was reasonably high for random effect terms on the log scale [55–57, 61]. The results were obtained using the WinBUGS software with two chains of MCMC chains and 100,000 posterior samplers were collected after the burn-in part of 100,000.

**Evaluation metrics**

To assess the model assumptions under different simulation conditions, five metrics were used to evaluate the



**Fig. 3** Plots of coefficients with lagged effect ( $\beta$ ) of two assumed situations used in the simulation study. Lag shown in weeks

performance of models: bias, root mean squared error (RMSE), credible interval (CrI) coverage probability, mean squared predictive error (MSPE) and model information criteria. The first measure was bias computed as the average difference between the simulated (true) mean and its estimate across the simulated datasets in each scenario. This measure is preferred to be near zero. To investigate the estimation uncertainty, we then also calculate the RMSE, summation of the variance of estimation with the bias squared, which was calculated as the squared root of the average squared deviation between the true and estimated means across the simulation replicates. The next metric was the coverage probability calculated as the percentage of the credible interval containing the generated value and preferred to have coverage similar to the pre-specified level. In this simulation study we pre-determined the credibility level at 95%. Then the mean squared predictive error was calculated as the squared root of the average squared difference between the simulated data and the predicted incidence from the posterior predictive distribution across the simulation replications. The posterior predictive distribution of  $y_{it}$  is in the form of  $p(y_{it}|\mathbf{y}) = \int_{\Theta} p(y_{it}|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$  where  $\mathbf{y}$  and  $\boldsymbol{\theta}$  denote all data and parameters used to fit the model. Finally, the deviance information criterion (DIC) was also used as the global metric for goodness of fit of models. For any sample primary parameter value  $\theta^g$  for the conditional likelihood, the deviance is  $D(\theta^g) = -2 \log f_{\theta^g|y}(y|\theta^g)$  and  $\bar{D}$  is the average deviance over the  $g$  posterior samplers. The effective number of parameters ( $pD$ ) is estimated as  $pD = \bar{D} - D(\bar{\theta})$ , and finally,  $DIC = \bar{D} + pD$ . The DIC was calculated under the assumption of single point estimate of the primary parameter of interest which may not be appropriate for mixture modeling as we assumed for zero inflation [29, 62]. Then an alternative is to compute  $pDr$  as the variance of deviance and define another measure variant as  $DICr = \bar{D} + pDr$ .

### Simulation results

Results of evaluation metrics are shown in Tables 1 and 2. For DLM and ZIPDLM, results were averaged over the spline basis orders. Figures 4-5 represent the estimated lagged effect curves of the true (simulated) coefficients with various simulation scenarios and parameter values. Overall, the distributed lag modeling appeared superior to the unconstrained counterpart, and zero-inflation models outperformed the models which did not account for excessive zeros. All zero-inflation models had similar estimations of zero proportions,  $\pi$ . For bias comparison, zero-inflation models had smaller bias while one-part models, GLM and DLM, had negative bias. This suggested that without properly accounting for excessive zeros the estimates can be negatively biased due to

the large number of zeros pulling the estimates towards zero. To investigate the variation in estimation, RMSEs were calculated under model assumptions. In general, the ZIP models showed smaller RMSEs than one-part models. Interestingly in situations with high collinearity, e.g.  $\rho = 0.8$ , the credible bands of DLM appeared to be slightly narrower than ZIPDLM. However, the RMSE does not include only the variance information but is summation of the variance of an estimate plus the square of its bias. Since DLM yielded bigger bias due to excessive zeros, RMSEs of DLM were larger than ZIPDLM.

From this evidence we should not only evaluate on point estimates but also should compare interval estimation. To assess the interval estimation, we compared models using the coverage probability. ZIPDLM has the nearest coverage probability to the pre-specified level of 0.95 whereas other models yielded relatively lower coverage probabilities. According to the simulation results in Tables 1 and 2, DLM particularly had the least coverage especially under high collinearity and excessive zeros. This might be resulting from the interval estimation of narrow credible intervals combined bias pulling towards zero which made the intervals miss the true values. To evaluate the goodness of fit, we applied the MSPE and information criteria to compare overall model fitting. Although all models had similar MSPEs, ZIPDLM had slightly smaller (better) MSPE. In addition, ZIPDLM also had the best (smallest) DIC and DICr. In general, the ZIPDLM model shows an improved the best result across assessment measures in the simulation study.

## Results

### Case study of lagged effects on sparse malaria incidence in an elimination setting along the Thailand-Myanmar border

In this section, weekly malaria incidence data of Takin 2016 were used to demonstrate the application of the proposed models in the previous section. As a hot spot of malaria incidence sub-districts in Tak along the Thai-Myanmar border were chosen to be a case study to show the performance of the models developed in the previous section. In addition, fine scale subnational data, in this case at sub-district level, have rarely been investigated in previous studies of climate and incidence. All cases had infection confirmed by microscopy (thick and thin films) and rapid diagnostic tests done at hospitals (mostly government hospitals) and/or malaria posts in the study area. Records were put into the national surveillance format and joint into the database of the Bureau of Vector Borne Disease at the Ministry of Public Health by the data management unit of each health facility. Clinical and demographic including gender, occupation, age,



**Table 1** Simulation result evaluation metrics from different models in simulation scenario 1

Measure	Model	$\pi = 0.4$		$\pi = 0.8$	
		$\rho = 0.3$	$\rho = 0.8$	$\rho = 0.3$	$\rho = 0.8$
Bias	ZIPGLM	0.037	0.055	-0.008	0.012
	ZIPDLM	-0.003	0.008	-0.004	-0.006
	GLM	-0.178	-0.438	-0.093	-0.067
	DLM	-0.091	-0.217	-0.071	-0.051
RMSE	ZIPGLM	0.791	0.881	0.508	0.574
	ZIPDLM	0.487	0.505	0.288	0.314
	GLM	0.942	0.959	0.833	0.875
	DLM	0.552	0.568	0.471	0.485
Crl Coverage	ZIPGLM	0.773	0.710	0.764	0.700
	ZIPDLM	0.933	0.799	0.927	0.740
	GLM	0.665	0.522	0.726	0.682
	DLM	0.607	0.484	0.710	0.599
DIC (pD)	ZIPGLM	1839.276 (36.599)	1994.511 (37.958)	3577.608 (61.961)	4056.092 (63.363)
	ZIPDLM	1838.029 (36.446)	1991.848 (35.182)	3576.794 (62.771)	4053.807 (61.442)
	GLM	1871.325 (58.264)	2025.734 (48.241)	3610.005 (69.376)	4078.526 (69.081)
	DLM	1868.994 (45.269)	2030.592 (45.277)	3629.755 (68.752)	4076.888 (65.247)
DICr (pDr)	ZIPGLM	1839.296 (36.579)	1998.464 (39.911)	3577.627 (61.978)	4059.542 (68.232)
	ZIPDLM	1838.321 (36.223)	1992.372 (35.182)	3576.157 (65.141)	4055.133 (62.772)
	GLM	1881.538 (58.264)	2035.515 (49.288)	3642.345 (81.376)	4085.638 (74.148)
	DLM	1874.314 (50.584)	2040.936 (54.709)	3637.313 (74.097)	4082.538 (72.898)
MSPE	ZIPGLM	81.268	160.155	143.496	332.312
	ZIPDLM	81.108	159.774	143.082	315.261
	GLM	82.093	164.001	144.612	315.224
	DLM	81.609	165.531	144.096	313.981
$\pi$	ZIPGLM	0.399	0.402	0.802	0.797
	ZIPDLM	0.401	0.401	0.802	0.797
	GLM	-	-	-	-
	DLM	-	-	-	-

residential location, disease onset and treatment date were collated from the surveillance system. Individual data were anonymized for their privacy protection.

The sub-district level population data were extracted from the online public database of the Bureau of Registration Administration. Geographical and administrative unit information were collated from the Ministry of Interior. The updated official geographical information system files at sub-district level were supplied by the Department of Local Administration. Data on average daily temperature, average daily relative humidity and total daily rainfall were considered in this study and collected from the Meteorological Department from five weather stations across the region during the study period. The weather variables were estimated using inverse distance weighted interpolation with tension spline in ArcGIS software version 10.3.1 and aggregated to weekly intervals at sub-district level as total weekly rainfall and average weekly temperature and relative humidity.

Fig. 6 shows the lag-coefficients between sub-district malaria incidence and climatic factors. The four models in the previous section were used to investigate the shapes of their relationships with the three environmental variables. The distributed lag models, both DLM and ZIPDLM, yielded smaller credible intervals and smoother mean lagged estimates of the coefficients for all weather factors. For humidity, the association started with a slightly negative effect for early lags and had a peak at around weeks 3–4 followed by a slowly declining trend. The relationship with rainfall began with a positive effect and then decreased over time. The association dropped to zero around lag 3 and was negative thereafter. The negative bias in one-part modeling similar to the simulation also occurred in the association with temperature. It could be seen that lagged estimates in GLM had a relatively large negative coefficient which might have resulted from the excessive zeros in our data. So, we focused on the ZIPDLM model for the temperature effect. The trend

**Table 2** Simulation result evaluation metrics from different models in simulation scenario 2

Measure	Model	$\pi = 0.4$		$\pi = 0.8$	
		$\rho = 0.3$	$\rho = 0.8$	$\rho = 0.3$	$\rho = 0.8$
Bias	ZIPGLM	0.131	-0.052	0.082	0.008
	ZIPDLM	-0.039	-0.021	-0.014	-0.002
	GLM	-0.157	-0.408	-0.119	-0.099
	DLM	-0.175	-0.418	-0.091	-0.081
RMSE	ZIPGLM	0.874	0.932	0.727	0.559
	ZIPDLM	0.522	0.608	0.388	0.291
	GLM	1.186	1.124	0.828	0.669
	DLM	0.619	0.702	0.434	0.328
Crl Coverage	ZIPGLM	0.712	0.700	0.750	0.740
	ZIPDLM	0.933	0.817	0.923	0.803
	GLM	0.632	0.580	0.755	0.695
	DLM	0.619	0.468	0.756	0.630
DIC (pD)	ZIPGLM	1720.445 (35.864)	2012.758 (38.732)	3635.735 (62.453)	3983.145 (69.391)
	ZIPDLM	1718.355 (34.256)	2011.769 (37.748)	3633.122 (60.932)	3969.928 (61.206)
	GLM	1762.11 (48.589)	2058.123 (50.145)	3660.459 (68.444)	4002.552 (74.235)
	DLM	1754.994 (44.321)	2053.159 (46.978)	3658.244 (64.442)	4000.879 (72.926)
DICr (pDr)	ZIPGLM	1723.112 (39.782)	2016.447 (42.552)	3638.445 (65.534)	3986.157 (70.763)
	ZIPDLM	1719.562 (35.834)	2012.224 (38.026)	3634.559 (61.952)	3970.251 (62.529)
	GLM	1764.965 (55.123)	2059.624 (53.998)	3672.428 (79.952)	4009.482 (75.992)
	DLM	1762.889 (52.216)	2051.223 (52.217)	3667.393 (73.591)	4005.938 (73.576)
MSPE	ZIPGLM	98.299	140.603	196.451	289.122
	ZIPDLM	97.002	140.102	196.112	288.112
	GLM	99.506	140.989	197.125	289.756
	DLM	98.954	140.559	196.478	288.422
$\pi$	ZIPGLM	0.3976	0.425	0.789	0.789
	ZIPDLM	0.3976	0.425	0.791	0.791
	GLM	-	-	-	-
	DLM	-	-	-	-

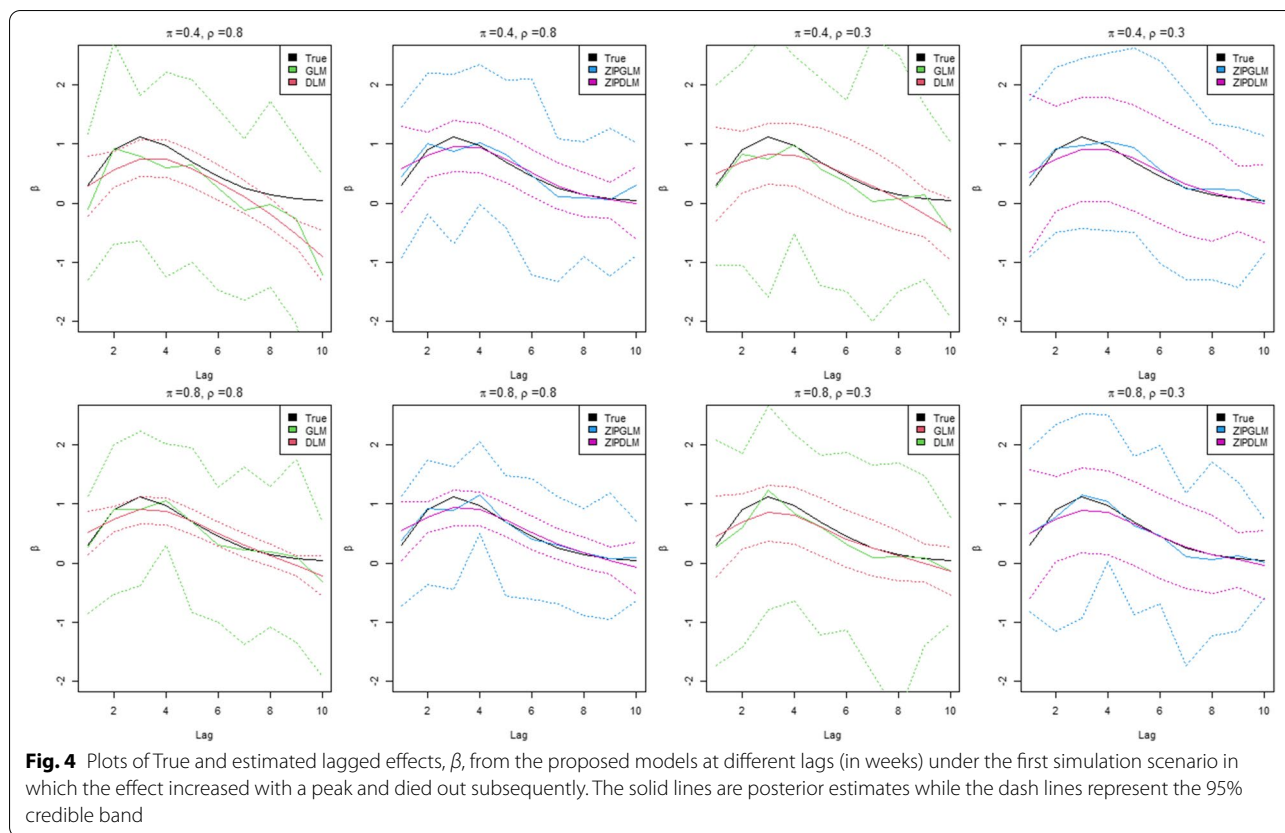
started negatively and then increased until around week 8 in which the curve remained afterwards. It seemed the proposed models, particularly ZIPDLM producing the best performance across evaluation measures, could be used to analyze the shape of association with malaria incidence useful for future disease control planning. However, additional issues were further discussed in the following section.

## Discussion

This study presents a new method to better quantify the association between climate and malaria over space and time which is suitable for near elimination settings. This can help better understand the impact of different factors on past spatiotemporal trends in incidence and to better predict future trends. It does so by simultaneously addresses the analytical challenges of accounting for the lag between changes in climate and changes in malaria transmission due to mosquito and parasite life cycles and

an excess of zero cases in the data encountered in low transmission settings. While count distributions, such as the negative binomial, can add flexibility by incorporating a separate heterogeneity parameter, these distributions do not always ensure adequate fit [43], especially in the case of excessive zeros. Without proper modeling, the high number of zeros tends to lower the mean estimates and neglect of excess zeroes can result in the biased estimation of parameters [44].

To address the analytical challenges, we developed a methodology which accommodates the zero inflation and delayed effects present in our data. In our simulation study, results showed that it was crucial to address both excessive zero cases and lagged nature of disease transmission. The models which did not accommodate for excessive zeros, i.e. GLM and DLM, yielded negatively biased estimates. This is important when planning for disease elimination because the bias can consequently lead to false recommendations and therefore potentially

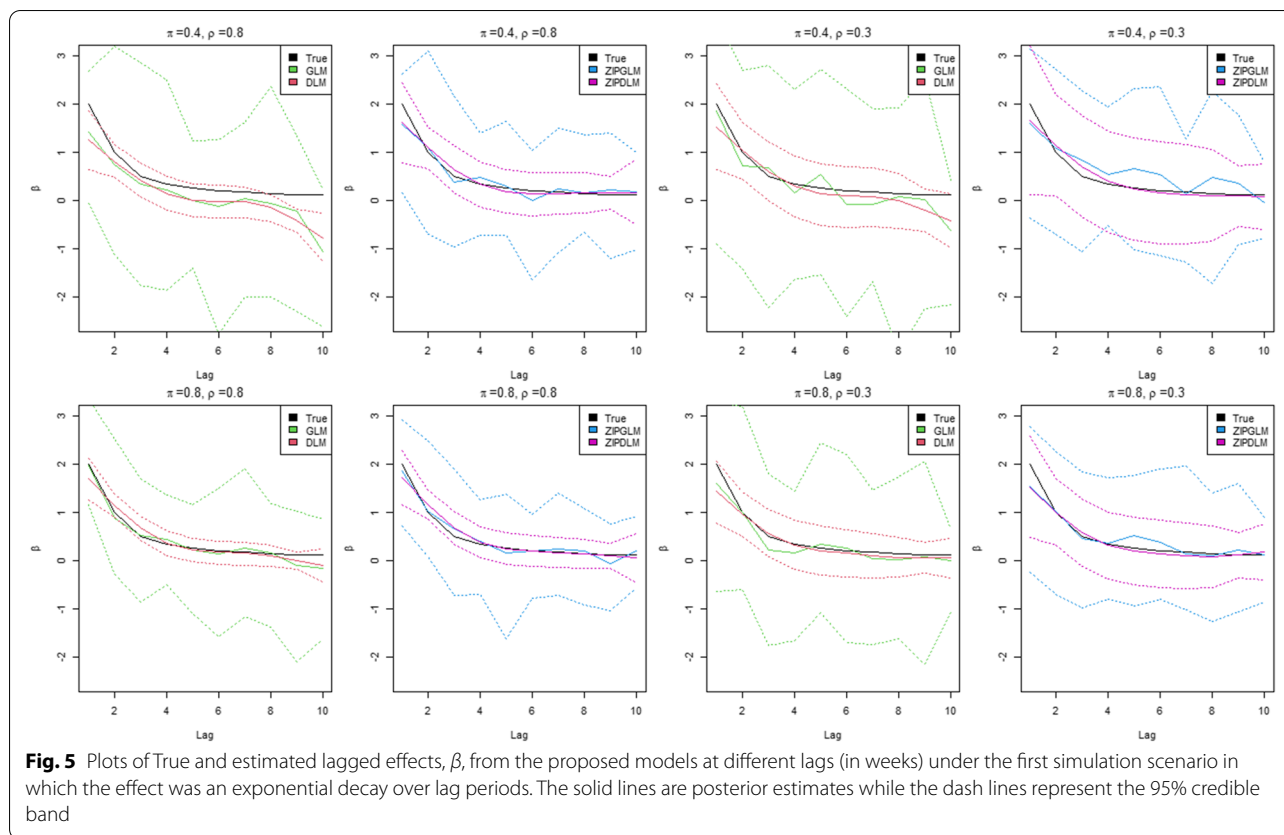


ineffective control activities. On the other hand, the unconstraint model produced high uncertain estimation due to collinearity between lags periods which could mask the significant association. In contrast, addressing only the lag effect could yield too narrow interval quantification. The model which accommodated both issues, ZIPDLM, appeared to have the overall best performance across evaluation metrics and simulation scenarios.

To demonstrate this with real data, we applied the developed framework to a case study of malaria incidence in Tak in which there were malaria hot spots. As suggested by previous studies there may be a temperature range critical for mosquito survivability [63, 64]. In addition, higher temperatures also accelerate multiplication of the *Plasmodium* parasites inside the vectors [14]. Since the peak period of malaria transmission in the GMS is usually in the rainy season which usually has a lower temperature, this could prolong the vector’s longevity in the region and hence the transmission could also occur at longer lags in the GMS. It has also been suggested that the optimal rainfall increases the adult mosquito’s size at early lags of 1 to 2 weeks [65] which was also similar to our finding. In our analysis, it was also found that the strongest associations of malaria with humidity occurred at lags 3–4 and declined thereafter. Increasing relative

humidity in an optimal lag period and range has a positive relationship with malaria incidence [66]. On the other hand, rainfall had a shorter lagged effect in our results with negative effect after a few weeks’ lag. Though increasing rainfall can make *Anopheles* larval habitats more receptive [67], flooding due simply to redundantly heavy rainfall, which often occurs in Thailand during the rainy season, may affect the vector’s life cycle. It has been shown that extreme weather can also impact the survival of vectors [68, 69] while light rainfall with suitable amounts of rain may be a more suitable condition for mosquito development.

There are several limitations in our study that should be acknowledged. We only considered three climatic factors (average humidity, total rainfall and average temperature) in association with malaria incidence due to data availability, although a wide range of other climate variables exist [70]. Other environmental variables than climate such as from remote sensing e.g. normalized difference vegetation index (NDVI) and surface water could also be incorporated [71]. Nonetheless, as the malaria parasite is transmitted from human to human via the bite of infected female mosquitoes of the genus *Anopheles* and the malaria case distribution and dynamics have been found to be closely related to environmental factors, it is sensible to include this nature



of the transmission process when available. Epidemiologically there is a lag between changes in environmental and entomological factors and malaria transmission due to both the life cycle of mosquitoes. Hence, we have developed the methodology to deal with the lagged time-varying distribution of the association which can partly account for the fluctuation of the environmental and entomological variates.

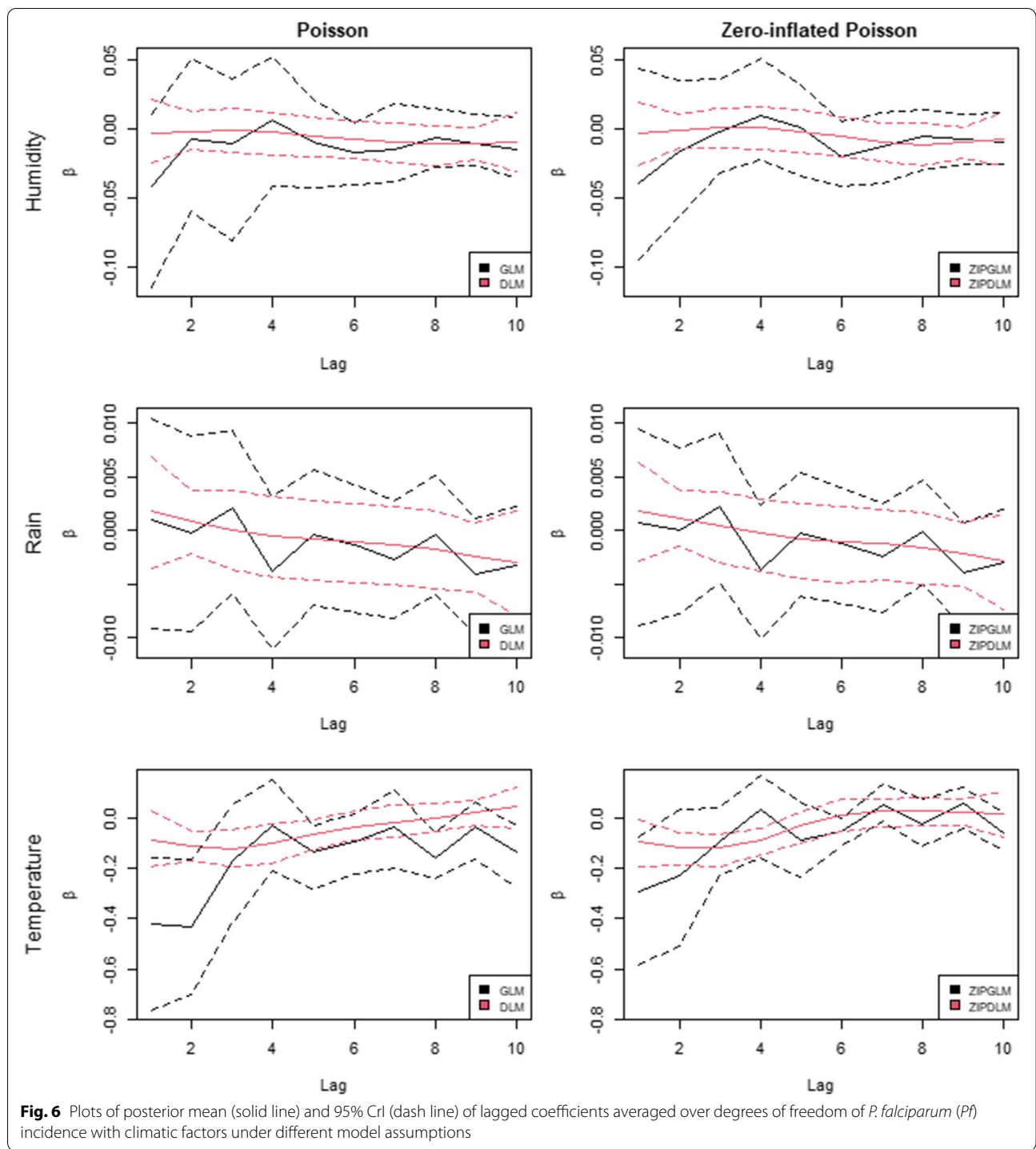
Malaria modeling is not an easy task due to the complex nature and difficulties in data collection. Various factors including entomological and population movement factors are important to investigate malaria transmission especially in the GMS region. In addition, the asymptomatic malaria may play important role in particular when malaria transmission comes to a very low level or very close to elimination in which the imported malaria cases may also contribute more than indigenous cases to malaria incidence. However, due to limited data and complex nature of malaria transmission, we could only apply secondary data from the national passive surveillance system in the study. Nonetheless, there is a need for future development to include those factors and the developed platform can also be utilized and extended to include those epidemiological and entomological variables as available to form a more complex relationship with malaria incidence. Therefore, it is important to

further consider confounding factors both at individual and population levels in future studies.

Even though there has been extensive research on factors that influence malaria transmission, the estimations that have been made in relation to climate in malaria endemic areas have been inconsistent across different regions [33]. Due to the complex relationship between malaria incidence, gaps in understanding still exist in the underlying processes of the linkage. Despite the limitations due to data availability the models developed here, particularly those accommodating excessive zero cases and lag effects, should provide a practical and useful step forwards in developing methods to examine the influence of different factors on malaria transmission as new data become available.

**Conclusions**

In this study, we have developed a methodology which accommodates zero inflation and lagged complexities of the relationship between malaria transmission and weather dynamics. The modeling assumptions were compared and discussed to identify the most appropriate approaches. We also demonstrated proposed models with real data using weekly malaria incidence data at a fine spatial scale. Because of the complexity of malaria transmission, gaps in knowledge still exist



in the underlying mechanisms. The developed methodology is promising and may have the potential to help better understand and improve estimates of past and predict future trends in malaria incidence. This could help policymakers to more effectively distribute resources and plan strategies for malaria elimination.

**Abbreviations**

GMS: Great Makong sub-region; DLM: Distributed Lag Model; GLM: Generalized Linear Model; ZIP: Zero-Inflation Poisson; ICAR: Intrinsic Conditional Autoregressive model; MSE: Mean Squared Error; MSPE: Mean Squared Predictive Error; DIC: Deviance Information Criterion; MCMC: *Markov chain Monte Carlo*; RMSE: Root Mean Squared Error; CrI: Credible Interval; NDVI: Normalized Difference Vegetation Index.

### Acknowledgements

We are grateful to the Nattawan Aunnarongrat who contributed her valuable time for data entry and editing composite data, Dr. Anchalee Jatapai for epidemiological interpretation, and Vilasinee Yuwaree who contributed to formatting and writing the ethical submission. We acknowledge the staff of the Bureau of Vector Borne Disease who were involved in management and the Department of Provincial Administration who supported and guided collection of the data.

### Authors' contributions

CR conceptualized the study, developed the statistical methodology, completed analyses, and drafted the manuscript. NE assisted with the epidemiological and climatic data. RJM and PS were responsible for revision and improvements of the manuscript and supervised the work. All authors have read and approved the final manuscript.

### Funding

This research was funded in part by the Bill and Melinda Gates Foundation (OPP1193472\_2018), the Faculty of Tropical Medicine, Mahidol University, and the Wellcome Trust [Grant number 106698/Z/14/Z]. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

### Availability of data and materials

The data that support the findings of this study were obtained from the Thai Bureau of Epidemiology, Ministry of Public Health, but restrictions apply to the availability of these data, which were used with permission for the current study, and are therefore not publicly available. However, data may be available from the authors upon reasonable request and with permission of the Thai Bureau of Epidemiology.

### Declarations

#### Ethics approval and consent to participate

This study protocol was approved by the Ethics Committee for Research in Human Subjects, Department of Disease Control, Ministry of Public Health, under the Royal Thai Government. The reference number is 1/60–001 and FWA number is 00013622. The aggregate data of this study were generated as part of the national routine surveillance system of communicable disease. Since the study was retrospective at the aggregated level in nature, the informed consent from subjects was waived by the Ethics Committee for Research in Human Subjects, Department of Disease Control, Ministry of Public Health, under the Royal Thai Government and the data were analyzed anonymously according to the guidelines of Mahidol University.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Department of Tropical Hygiene, Faculty of Tropical Medicine, Mahidol University, Ratchathewi, Bangkok 10400, Thailand. <sup>2</sup>Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand. <sup>3</sup>Division of Vector Borne Diseases, Department of Disease Control, Ministry of Public Health, Nonthaburi, Thailand. <sup>4</sup>Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, MA, USA. <sup>5</sup>Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. <sup>6</sup>The Open University, Milton Keynes, UK.

Received: 31 May 2021 Accepted: 22 November 2021

Published online: 20 December 2021

### References

- Kumar A, Valecha N, Jain T, Dash AP. Burden of malaria in India: retrospective and prospective view. *The American journal of tropical medicine and hygiene*. 2007;77(6\_Suppl):69–78.
- Zhou G, Sirichaisinthop J, Sattabongkot J, Jones J, Bjørnstad ON, Yan G, et al. Spatio-temporal distribution of plasmodium falciparum and P. vivax malaria in Thailand. *The American journal of tropical medicine and hygiene*. 2005;72(3):256–62.
- World Health O. World malaria report. 20 years of global progress and challenges. Geneva: World Health Organization. 2020;2020:2020.
- Carrara VI, Sirilak S, Thonglairuam J, Rojanavatsirivet C, Proux S, Gilbos V, et al. Deployment of early diagnosis and mefloquine-artesunate treatment of falciparum malaria in Thailand: the Tak malaria initiative. *PLoS Med*. 2006;3(6):e183.
- Parker DM, Landier J, Thu AM, Lwin KM, Delmas G, Nosten FH, et al. Scale up of a plasmodium falciparum elimination program and surveillance system in Kayin State, Myanmar. 2017;2.
- Zhou G, Minakawa N, Githeko AK, Yan G. Association between climate variability and malaria epidemics in the east African highlands. *Proc Natl Acad Sci*. 2004;101(8):2375–80.
- Gunda R, Chimbari MJ, Shamu S, Sartorius B, Mukaratirwa SJMJ. Malaria incidence trends and their association with climatic variables in rural Gwanda, Zimbabwe, 2005–2015. 2017;16(1):1–13.
- Hundessa S, Williams G, Li S, Guo J, Zhang W, Guo Y. The weekly associations between climatic factors and plasmodium vivax and plasmodium falciparum malaria in China, 2005–2014. *Trans R Soc Trop Med Hyg*. 2017;111(5):211–9.
- Xiao D, Long Y, Wang S, Fang L, Xu D, Wang G, et al. Spatiotemporal distribution of malaria and the association between its epidemic and climate factors in Hainan, China. *Malaria journal*. 2010;9(1):185.
- Kim Y, Ratnam J, Doi T, Morioka Y, Behera S, Tsuzuki A, et al. Malaria predictions based on seasonal climate forecasts in South Africa: a time series distributed lag nonlinear model. *Sci Rep*. 2019;9(1):1–10.
- Kumar P, Vatsa R, Sarthi PP, Kumar M, Gangare V. Modeling an association between malaria cases and climate variables for Keonjhar district of Odisha, India: a Bayesian approach. *J Parasit Dis*. 2020;44(2):319–31.
- Grover-Kopec EK, Blumenthal MB, Ceccato P, Dinku T, Omumbo JA, Connor SJ. Web-based climate information resources for malaria control in Africa. *Malar J*. 2006;5(1):1–9.
- Wardrop NA, Barnett AG, Atkinson J-A, Clements AC. Plasmodium vivax malaria incidence over time and its association with temperature and rainfall in four counties of Yunnan Province. *China Malaria journal*. 2013;12(1):452.
- Loiseau C, Harrigan RJ, Bichet C, Julliard R, Garnier S, Lendvai AZ, et al. Predictions of avian plasmodium expansion under climate change. *Sci Rep*. 2013;3(1):1–6.
- Sena LD, Deressa WA, Ali AA. Analysis of trend of malaria prevalence in south-West Ethiopia: a retrospective comparative study. *Malar J*. 2014;13(1):188.
- Zhao X, Chen F, Feng Z, Li X, Zhou X-H. The temporal lagged association between meteorological factors and malaria in 30 counties in south-West China: a multilevel distributed lag non-linear analysis. *Malar J*. 2014;13(1):57.
- Hundessa S, Williams G, Li S, Li Liu D, Cao W, Ren H, et al. Projecting potential spatial and temporal changes in the distribution of Plasmodium vivax and Plasmodium falciparum malaria in China with climate change. 2018;627:1285–93.
- Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. *Stat Med*. 2010;29(21):2224–34.
- Gasparrini A. Distributed lag linear and non-linear models in R: the package dlnm. *J Stat Softw*. 2011;43(8):1.
- Gasparrini A. Modeling exposure–lag–response associations with distributed lag non-linear models. *Stat Med*. 2014;33(5):881–99.
- Schwartz J. The distributed lag between air pollution and daily deaths. *Epidemiology*. 2000;11(3):320–6.
- Zanobetti A, Wand M, Schwartz J, Ryan L. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*. 2000;1(3):279–92.
- Welty LJ, Zeger SL. Are the acute effects of particulate matter on mortality in the National Morbidity, mortality, and air pollution study the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. *Am J Epidemiol*. 2005;162(1):80–8.
- Wu Y, Qiao Z, Wang N, Yu H, Feng Z, Li X, et al. Describing interaction effect between lagged rainfalls on malaria: an epidemiological study in south–West China. *Malar J*. 2017;16(1):1–10.

25. Warren JL, Luben TJ, Chang HH. A spatially varying distributed lag model with application to an air pollution and term low birth weight study. *J R Stat Soc: Ser C: Appl Stat.* 2020;69(3):681–96.
26. Warren J, Fuentes M, Herring A, Langlois P. Spatial-temporal modeling of the association between air pollution exposure and preterm birth: identifying critical windows of exposure. *Biometrics.* 2012;68(4):1157–67.
27. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data: CRC press; 2014.
28. Lawson AB, Rotejanaprasert C. Childhood brain cancer in Florida: a Bayesian clustering approach. *Statistics and Public Policy.* 2014;1(1):99–107.
29. Thway AM, Rotejanaprasert C, Sattabongkot J, Lawawirojwong S, Thi A, Hlaing TM, et al. Bayesian spatiotemporal analysis of malaria infection along an international border: Hlaingbwe township in Myanmar and Thasong-Yang District in Thailand. *Malar J.* 2018;17(1):428.
30. Neelon B, Chang HH, Ling Q, Hastings NS. Spatiotemporal hurdle models for zero-inflated count data: exploring trends in emergency department visits. *Stat Methods Med Res.* 2016;25(6):2558–76.
31. Arab A. Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *Int J Environ Res Public Health.* 2015;12(9):10536–48.
32. Abiodun GJ, Makinde OS, Adeola AM, Njabo KY, Witbooi PJ, Djidjou-Demasse R, et al. A dynamical and zero-inflated negative binomial regression modelling of malaria incidence in Limpopo Province, South Africa. *Int J Environ Res Public Health.* 2019;16(11):2000.
33. Teklehaimanot HD, Lipsitch M, Teklehaimanot A, Schwartz JJMj. Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. 2004;3(1):41.
34. Srimath-Tirumula-Peddinti RCPK, Neelapu NRR, Sidagam NJPO. Association of climatic variability, vector population and malarial disease in district of Visakhapatnam. India: a modeling and prediction analysis. 2015;10(6):e0128377.
35. Ross A, Koepfli C, Schoepfli S, Timinao L, Siba P, Smith T, et al. The incidence and differential seasonal patterns of plasmodium vivax primary infections and relapses in a cohort of children in Papua New Guinea. *PLoS Negl Trop Dis.* 2016;10(5):e0004582.
36. Rotejanaprasert C, Lee D, Ekapirat N, Sudathip P, Maude RJ. Spatiotemporal distributed lag modelling of multiple plasmodium species in a malaria elimination setting. *Stat Methods Med Res.* 2021;30(1):22–34.
37. White MT, Karl S, Koepfli C, Longley RJ, Hofmann NE, Wampfler R, et al. Plasmodium vivax and plasmodium falciparum infection dynamics: re-infections, recrudescences and relapses. *Malar J.* 2018;17(1):1–15.
38. Pascutto C, Wakefield J, Best N, Richardson S, Bernardinelli L, Staines A, et al. Statistical issues in the analysis of disease mapping data 2000;19(17–18):2493–2519.
39. Lawson AB, Banerjee S, Haining RP, Ugarte MD. Handbook of Spatial Epidemiology: CRC Press; 2016.
40. Best N, Richardson S, Thomson AJSmimr. A comparison of Bayesian spatial models for disease mapping. 2005;14(1):35–59.
41. Waller LA, Gotway CA. Applied spatial statistics for public health data: John Wiley & Sons; 2004.
42. Lawson AB. Bayesian disease mapping: hierarchical modeling in spatial epidemiology: CRC press; 2013.
43. Neelon B, O'Malley AJ, Smith VA. Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Stat Med.* 2016;35(27):5070–93.
44. Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. *Stat Med.* 2002;21(10):1461–9.
45. Corpas-Burgos F, García-Donato G, Martínez-Beneito MA. Some findings on zero-inflated and hurdle poisson models for disease mapping. *Stat Med.* 2018;37(23):3325–37.
46. Albert JM, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Stat Methods Med Res.* 2014;23(3):257–78.
47. Preisser JS, Stamm JW, Long DL, Kincade ME. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Res.* 2012;46(4):413–23.
48. Almon S. The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society.* 1965:178–96.
49. Hajat S, Armstrong BG, Gouveia N, Wilkinson P. Mortality displacement of heat-related deaths: a comparison of Delhi, Sao Paulo, and London. *Epidemiology.* 2005;613–20.
50. Zanobetti A, Schwartz J, Samoli E, Gryparis A, Touloumi G, Atkinson R, et al. The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement. *Epidemiology.* 2002;13(1):87–93.
51. Rondeau V, Berhane K, Thomas DC. A three-level model for binary time-series data: the effects of air pollution on school absences in the Southern California Children's health study. *Stat Med.* 2005;24(7):1103–15.
52. Rotejanaprasert C, Ekapirat N, Areechokchai D, Maude RJ. Bayesian spatiotemporal modeling with sliding windows to correct reporting delays for real-time dengue surveillance in Thailand. *Int J Health Geogr.* 2020;19(1):1–13.
53. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math.* 1991;43(1):1–20.
54. Rotejanaprasert C, Ekapirat N, Areechokchai D, Maude RJ. Bayesian spatiotemporal modeling with sliding windows to correct reporting delays for real-time dengue surveillance in Thailand. *Int J Health Geogr.* 2020;19(1):4.
55. Rotejanaprasert C, Lawson AB, Iamsirithaworn S. Spatiotemporal multi-disease transmission dynamic measure for emerging diseases: an application to dengue and zika integrated surveillance in Thailand. *BMC Med Res Methodol.* 2019;19(1):200.
56. Rotejanaprasert C, Lawson A. Bayesian prospective detection of small area health anomalies using Kullback–Leibler divergence. *Stat Methods Med Res.* 2018;27(4):1076–87.
57. Rotejanaprasert C, Lawson AB. A bayesian quantile modeling for spatiotemporal relative risk: an application to adverse risk detection of respiratory diseases in South Carolina, USA. *Int J Environ Res Public Health.* 2018;15(9):2042.
58. Graham MHJE. Confronting multicollinearity in ecological multiple regression. 2003;84(11):2809–15.
59. De Marco P, Nóbrega CCJPO. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation 2018;13(9):e0202403.
60. Mateo RG, Vanderpoorten A, Muñoz J, Laenen B, Désamoré AJPO. Modeling species distributions from heterogeneous data for the biogeographic regionalization of the European bryophyte flora 2013;8(2):e55648.
61. Carroll R, Lawson AB, Faes C, Kirby RS, Aregay M, Watjoo KJSmimr. Spatially-dependent Bayesian model selection for disease mapping 2018;27(1):250–268.
62. Celeux G, Forbes F, Robert CP, Titterton DM. Deviance information criterion for missing data models. *Bayesian Anal.* 2006;1(4):651–73.
63. Bayoh MN, Lindsay SWJM, entomology v. Temperature-related duration of aquatic stages of the Afrotropical malaria vector mosquito *Anopheles gambiae* in the laboratory 2004;18(2):174–179.
64. Christiansen-Jucht CD, Parham PE, Saddler A, Koella JC, Basáñez M-GJP, vectors. Larval and adult environmental temperatures influence the adult reproductive traits of *Anopheles gambiae* ss 2015;8(1):456.
65. Kitthawee S, Edman J, Upatham E. Anopheles dirus size and fecundity: relationship to larval density and protein accumulation. *The Southeast Asian journal of tropical medicine and public health.* 1992;23(1):128–31.
66. Li T, Yang Z, Wang M. Temperature, relative humidity and sunshine may be the effective predictors for occurrence of malaria in Guangzhou, southern China, 2006–2012. *Parasit Vectors.* 2013;6(1):155.
67. McMichael AJ. Globalization, climate change, and human health. *N Engl J Med.* 2013;368(14):1335–43.
68. Egbendewe-Mondzozo A, Musumba M, McCarl BA, Wu XJJoER, health P. Climate change and vector-borne diseases: an economic impact analysis of malaria in Africa 2011;8(3):913–930.
69. Agosto FBJC. Optimal control and temperature variations of malaria transmission. *Dynamics.* 2020;2020.
70. The Global Observing System for Climate: Implementation Needs: World Meteorological Organization (WMO); 2016 [cited 2021 11 April]. Available from: [https://library.wmo.int/index.php?lvl=notice\\_display&id=19838#\\_YHKL7j8xWbL](https://library.wmo.int/index.php?lvl=notice_display&id=19838#_YHKL7j8xWbL).
71. Zinszer K, Verma AD, Charland K, Brewer TF, Brownstein JS, Sun Z, et al. A scoping review of malaria forecasting: past work and future directions. 2012;2(6).

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.