Supplemental Information A: Further investigations of GAM and SPDE-INLA performance

To better understand the reasons behind the poor performance of two of the best models (GAM and SPDE-INLA) in some countries, we further analyzed the cross-validation results and performed simulations. We postulate that GAM is prone to generate poor predictions close to the edge of the countries, resulting in its poor performance in certain countries. Since our spatio-temporal Gaussian process model uses "separable" covariance matrix, the model assumes that spatial correlation in the two time points remain the same. We postulate that violation of this assumption can cause this model formulation to perform worse under spatio-temporal setting when compared to spatial setting.

We focus on comparing the two models of interest in this section, GAM and SPDE-INLA, instead of comparing all five models for conciseness.

1. Impact of boundary issues on GAM performance

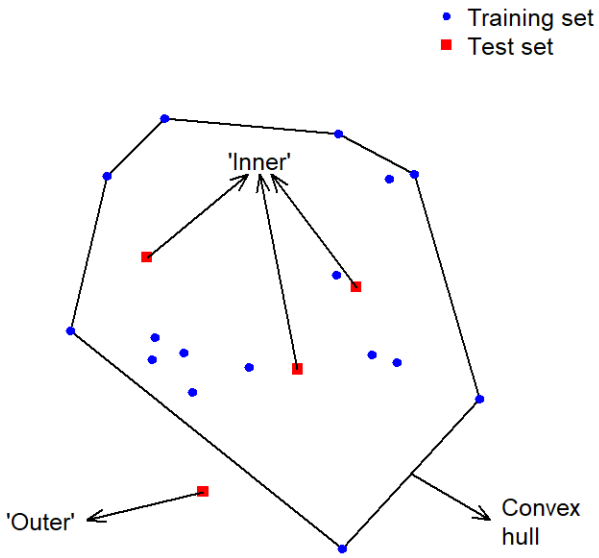
To determine if boundary issues exist for GAM's and explain the observed model performance results for GAM, we restrict our investigation of boundary issue to the spatial setting only, i.e. only the present data is used.

1.1. Analysis based on observed data

Using our ten-fold cross validation results, we first identified the sampling clusters that are "outside" of the training dataset. For each fold of cross validation, we created a convex hull based on the training set's coordinates. A sampling cluster in the test set that is outside of the convex hull is categorized as an "outer" cluster, whereas those within the convex hull are "inner" clusters (Fig. 1). For each country, about five to seven percent of the sampling clusters belong to the "outer" group. We compared the MAE and per-person log-likelihood of the outer vs inner group for each model in each country. If

GAM's boundary issue is prominent, we expect that the difference in the metrics between the two group is consistently larger in GAM than in other models.

Figure 1: "Inner" and "Outer" locations as defined in our convex hull analysis based on existing cross validation results.



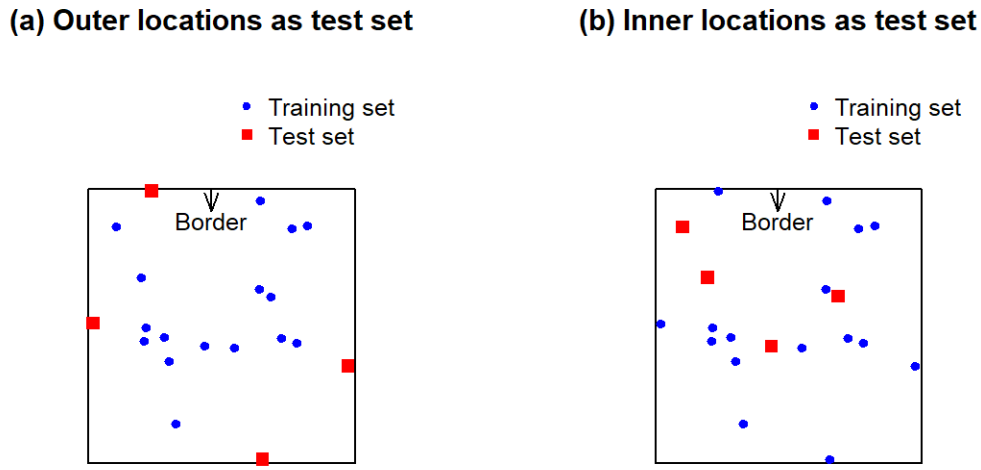
1.2. Analysis based on simulated data

We also used simulated data to further verify the boundary issue in GAM. We used existing data as template to generate simulated datasets. We retained the sampling locations, number of sampled children and four geospatial covariates associated with them (Population, Nightime Light, EVI and Temperature) in the existing datasets. We used a stationary Gaussian Process model (as outline in main text) to simulate the underlying probability of infections in each of these locations. For the covariance matrix (Σ), we set $\sigma^2 = 1$ and $\kappa = 1/20$. For each country, we randomly generated ten datasets. For each dataset, α and β were first randomly drawn (β was drawn from $N_5(0, I)$), and then the malaria

status of each individual is reassigned randomly based on the newly simulated probabilities.

To evaluate each model's predictive performance at the boundary of a country, two types of cross validations were conducted for each simulated dataset. First, we designated 10% of the sampling locations that are closest to the border of the country as test set, and we train each model using the remaining "inner" 90% of the data. We used the trained models to predict the probability of malaria infections and calculated the logLik and MAE using the same formula described in previous section. For the second set of cross validations, we randomly selected 10% of the "inner" sampling locations as the test set (See Fig. 2). Likewise, we trained the models using the remaining 90% and calculated the performance metrics. Thus for each country, we obtained 20 performance metrics, 10 of which were from the border or "outer" clusters and 10 from the "inner" clusters. We compared the mean of the metrics between the two groups and among the models. Similarly, we expect the difference in the prediction accuracies between the outer and inner clusters to be larger in GAM than in other models if GAM's boundary problem is severe.

Figure 2: In our simulation study to examine the boundary issue, we compare the models out-of-sample predictive performance between (a) using "Outer" locations (closest points to the border) and (b) using "Inner" locations (randomly selected points that are not the closest to the border).



2. Impact of temporal changes in spatial correlation on SPDE-INLA performance

2.1. Analysis based on observed data

To test our hypothesis that poorer performance of SPDE-INLA under spatio-temporal setting vs spatial setting can be attributed to distinctive spatial correlation at two time points, we analyzed the spatial dependencies in the present and past data separately. For each country, we fit the spatial SPDE-INLA to the present dataset and obtain the posterior mean of the (inverse) range parameter, κ_{pres}. Larger κ indicates weaker spatial correlation and vice versa. Following a similar procedure, we also obtained the posterior mean of the range parameter in the past dataset, κ_{past}. We examined the relationship between prediction

accuracy of each model under the spatio-temporal setting and the ratio of the range parameter ($\kappa_{past}/\kappa_{pres}$).

2.2. Analysis based on simulated data

We conducted a simulation exercise to find additional evidence. We focus on generating simulated datatsets using the existing Burkina Faso data as template. Similar to Section 1, we retained all information in both past and present data and modified the malaria prevalences in each sampling cluster based on two independent Gaussian Process model with four geospatial covariates, i.e. the Gaussian Process in the past dataset is uncorrelated with the Gaussian process in the present dataset. For the spatial Matern covariance function, we let the range parameter to vary according to the two time points. We created three simulated datasets for each ratio of range parameter ($\kappa_{past}/\kappa_{pres}$). To this end, we fixed κ_{pres} to 1/64 and varied κ_{past} so that the ratio $\kappa_{past}/\kappa_{pres}$ would be equal to 1, 2, 4, 6, or 8. For each dataset we fixed $\sigma^2 = 1$, randomly drew α and β, and reassigned the malaria status of each individual based on the newly simulated probabilities.

For each simulated dataset, we evaluated each spatio-temporal model's performance using ten-fold cross-validation. If the prediction accuracy of SPDE-INLA decreases with respect to other models as the ratio of range parameter increases, it would support our postulation that changing spatial-dependency between two time points affects our SPDE-INLA more than other models.