Landscape disease ecology journal club

Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: A case study predicting malaria risk in Mozambique

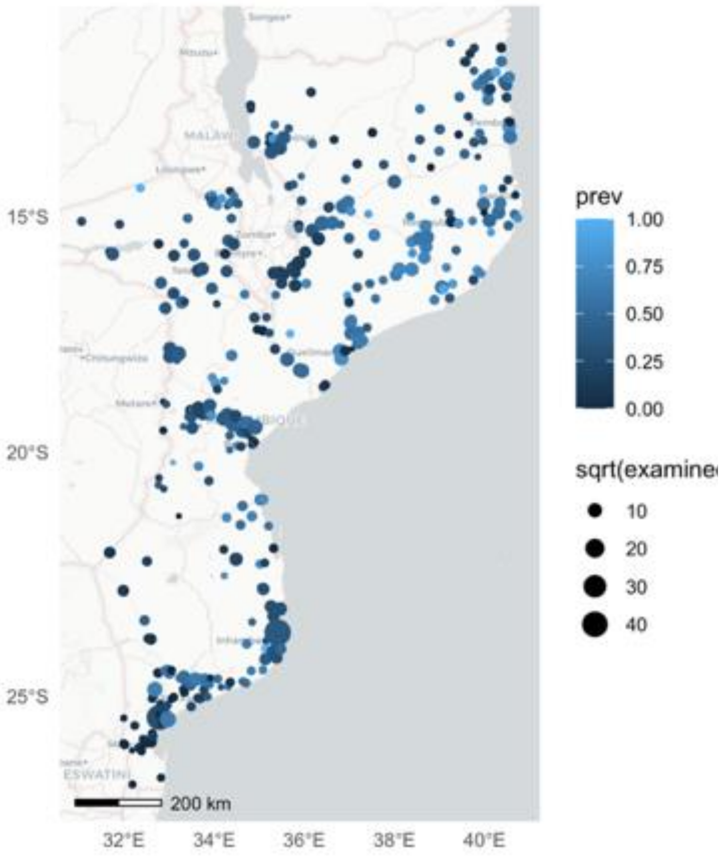
Moraga *et al.* 2021

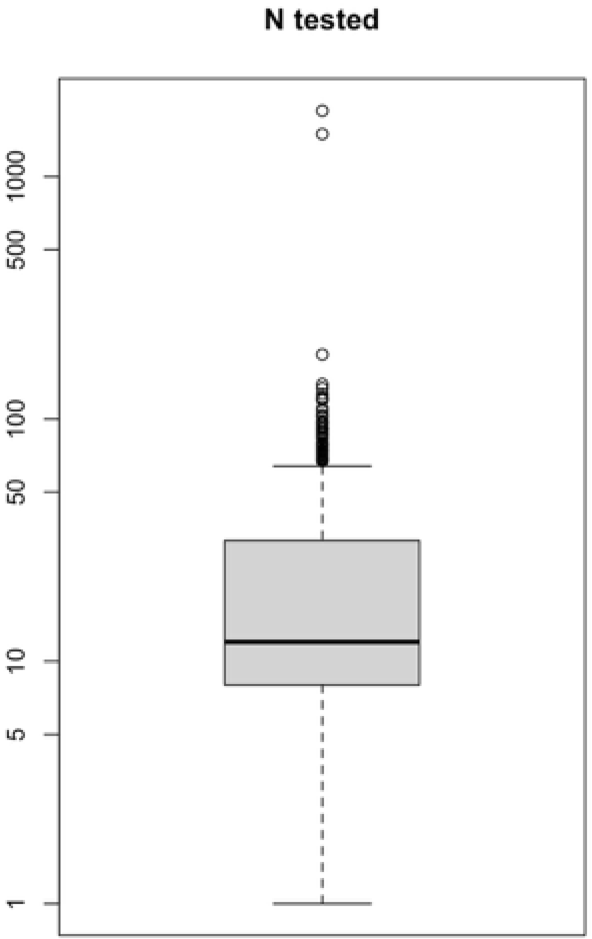
https://doi.org/10.1016/j.sste.2021.100440

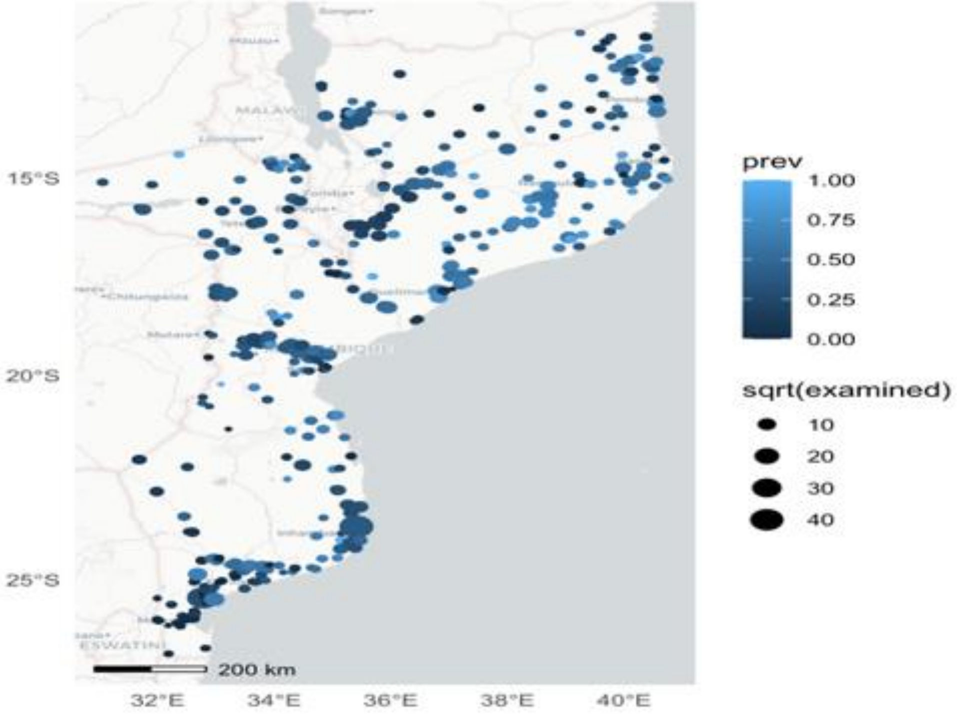
What is INLA (integrated nested Laplace approximation)?

- “INLA is a method for **approximate Bayesian inference** in latent Gaussian models which are a useful subclass of structured additive regression models.”
- “INLA has gained attention as a suitable alternative to MCMC particularly when dealing with big data, and **in applications involving spatial data where MCMC methods do not work as efficiently** with the computing resources available.”
- “Moreover, INLA can be used in combination with the stochastic partial differential equation (SPDE) approach to analyse **spatial point data**.”

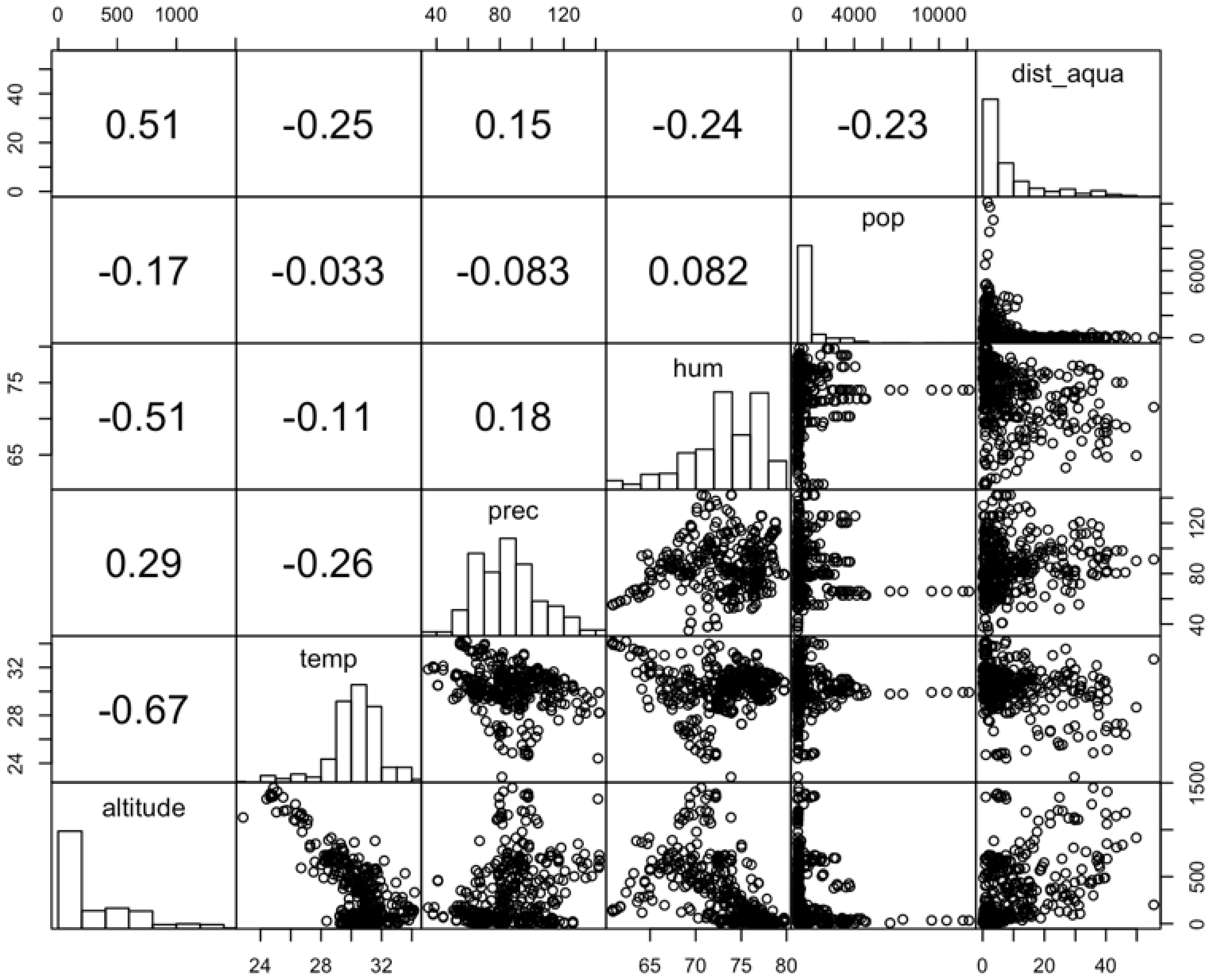
Data: 447 malaria prevalence surveys in Mozambique, from the year 2000-…



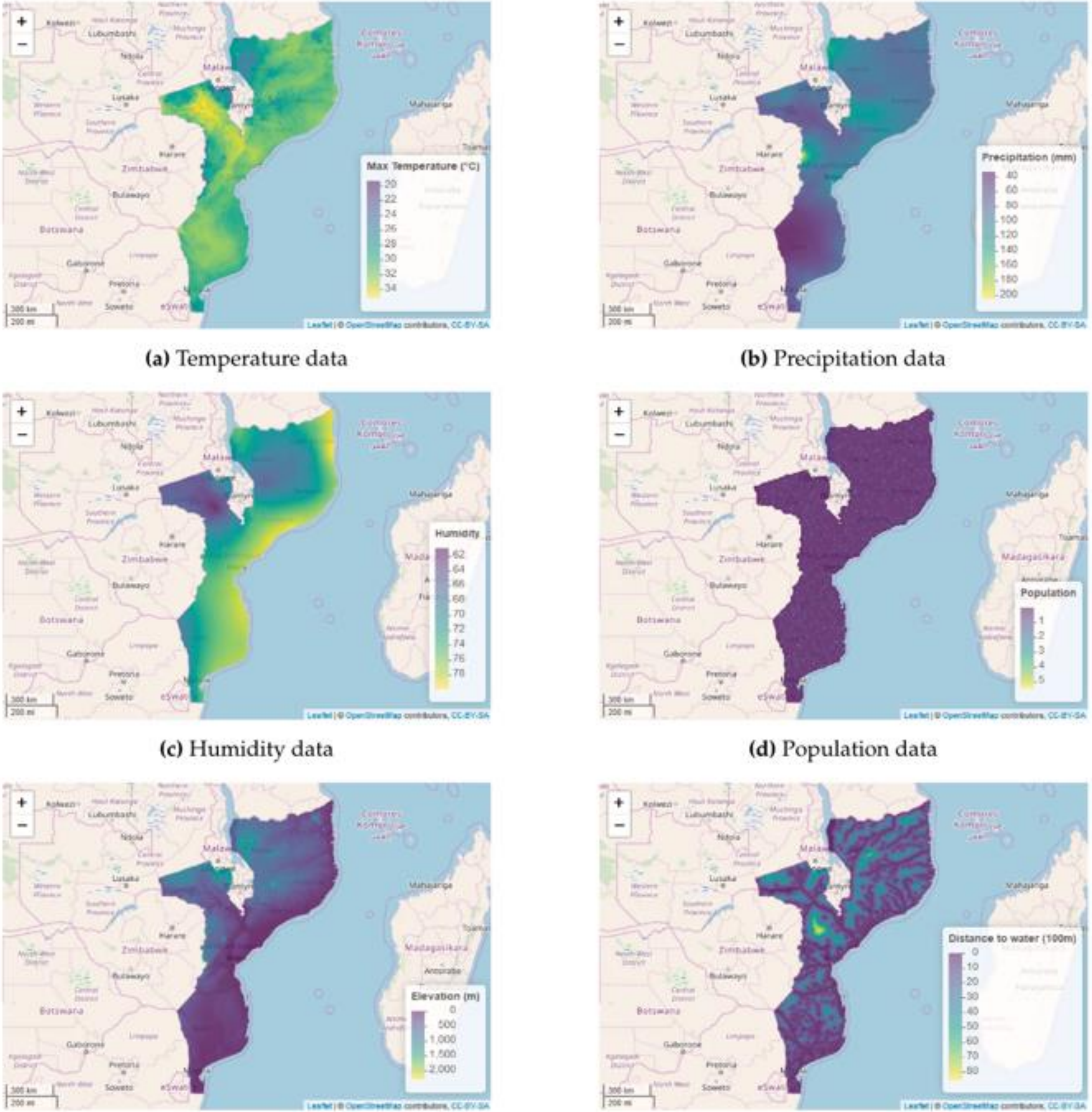




Data: covariates



Data: covariates



(a) Temperature data

(b) Precipitation data

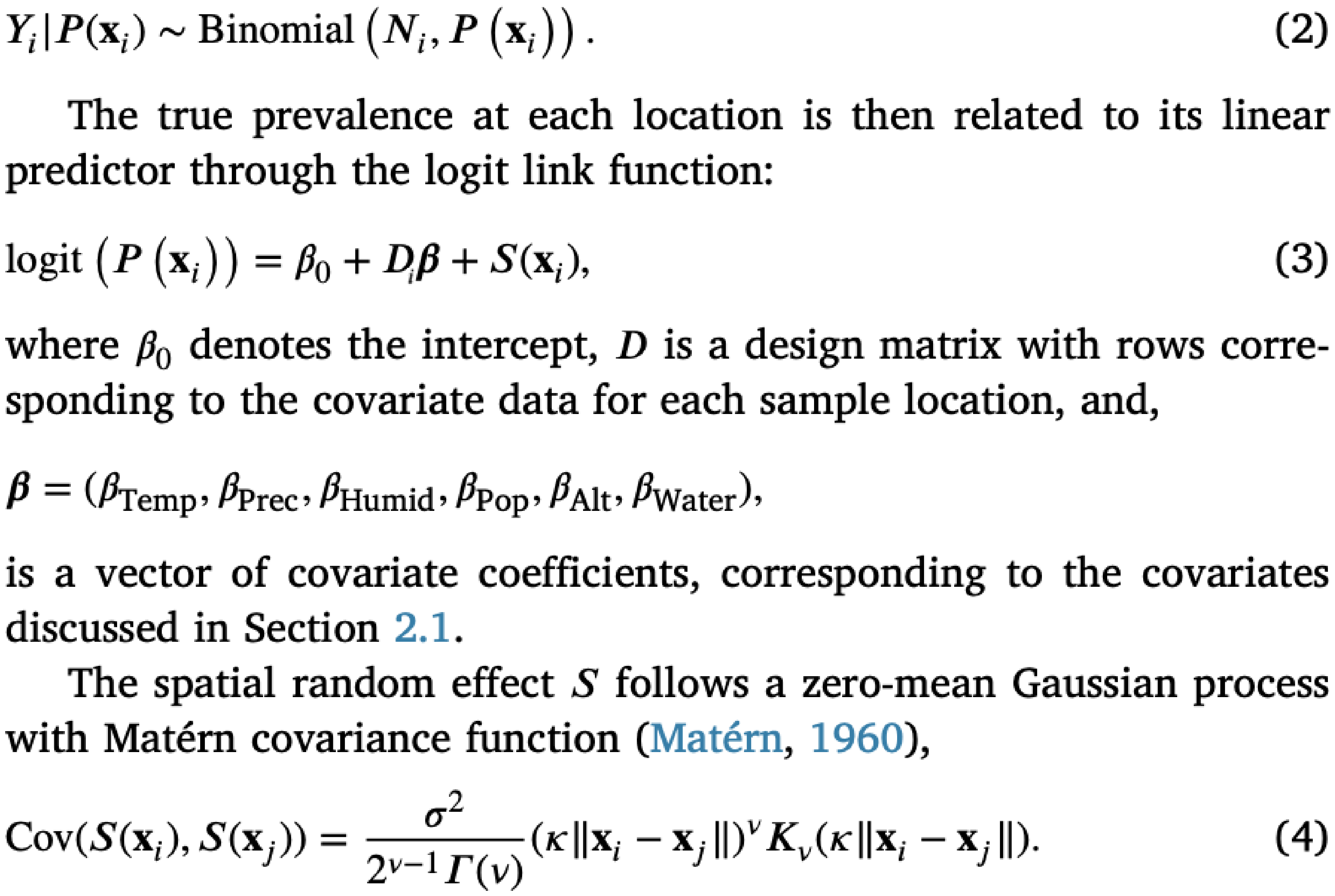
(c) Humidity data

(d) Population data

(e) Altitude data

(f) Distance of nearest inland water sources

The model



$$Y_i | P(\mathbf{x}_i) \sim \text{Binomial}\left(N_i, P\left(\mathbf{x}_i\right)\right). \tag{2}$$

The true prevalence at each location is then related to its linear predictor through the logit link function:

$$\text{logit}\left(P\left(\mathbf{x}_i\right)\right) = \beta_0 + D_i\boldsymbol{\beta} + S(\mathbf{x}_i), \tag{3}$$

where β_0 denotes the intercept, D is a design matrix with rows corresponding to the covariate data for each sample location, and,

$$\boldsymbol{\beta} = (\beta_{\text{Temp}}, \beta_{\text{Prec}}, \beta_{\text{Humid}}, \beta_{\text{Pop}}, \beta_{\text{Alt}}, \beta_{\text{Water}}),$$

is a vector of covariate coefficients, corresponding to the covariates discussed in Section 2.1.

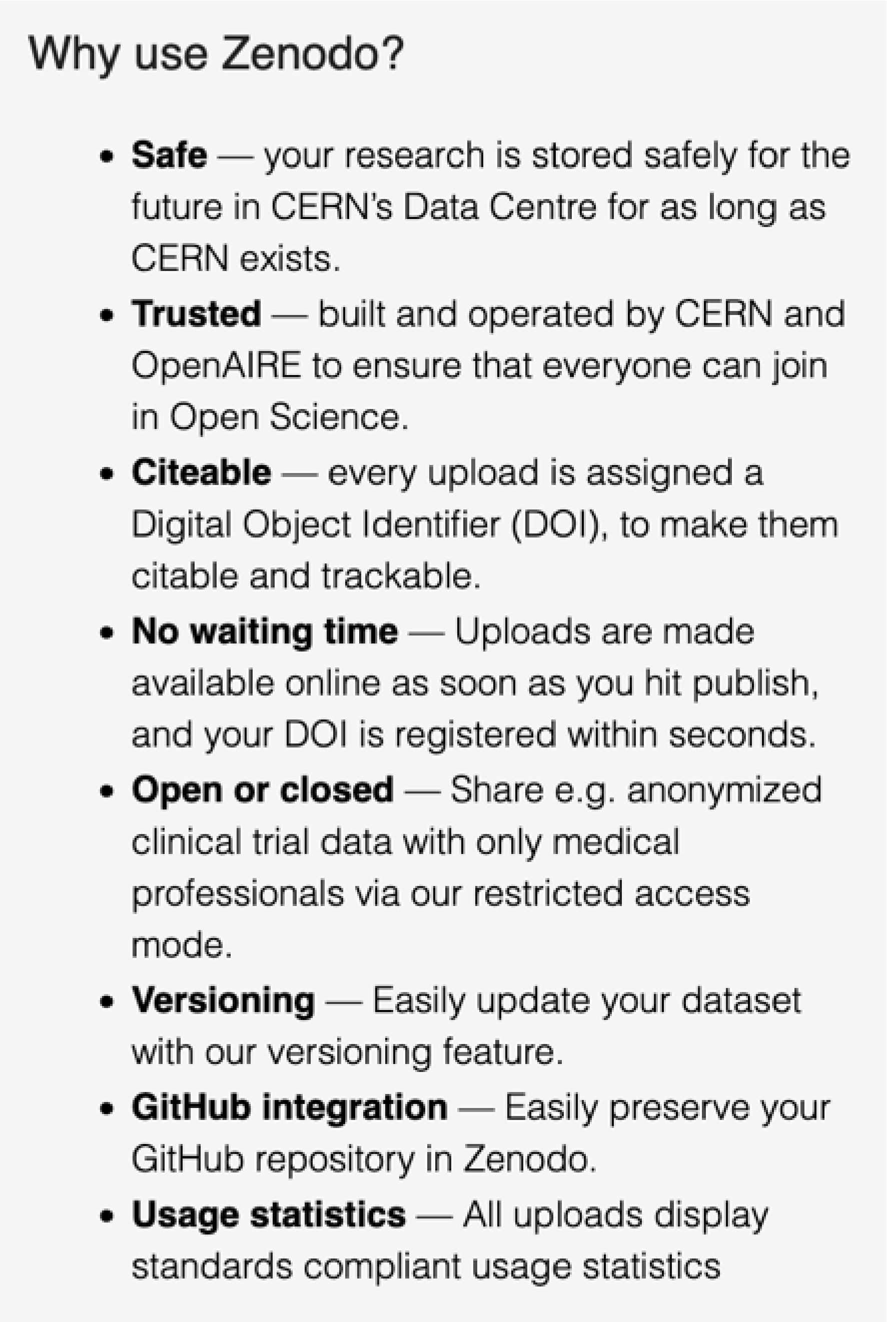
The spatial random effect S follows a zero-mean Gaussian process with Matérn covariance function (Matérn, 1960),

$$\text{Cov}(S(\mathbf{x}_i), S(\mathbf{x}_j)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa\|\mathbf{x}_i - \mathbf{x}_j\|)^{\nu} K_{\nu}(\kappa\|\mathbf{x}_i - \mathbf{x}_j\|). \tag{4}$$

Fitting the model with R-INLA

- All the data and code is available as an R project
  - https://github.com/Paula-Moraga/spatial-mod
- ...except it isn't

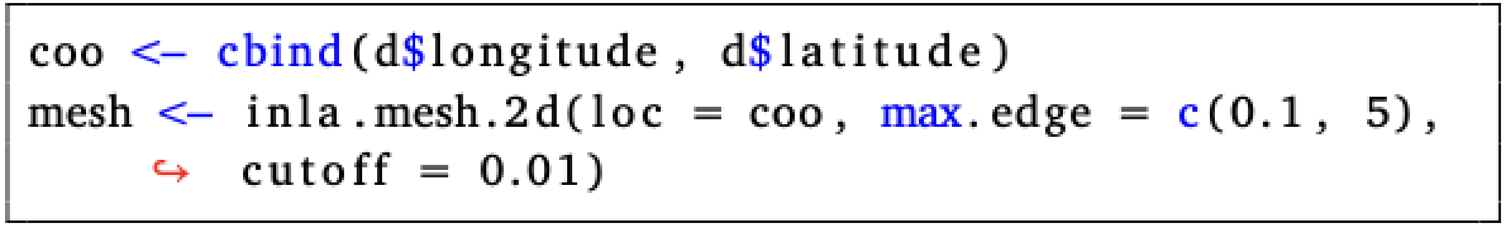
- ...oh but it is
  - https://github.com/ncespedesc/spatial-mode
- ...and I've forked it
  - https://github.com/pcdjohnson/spatial-model



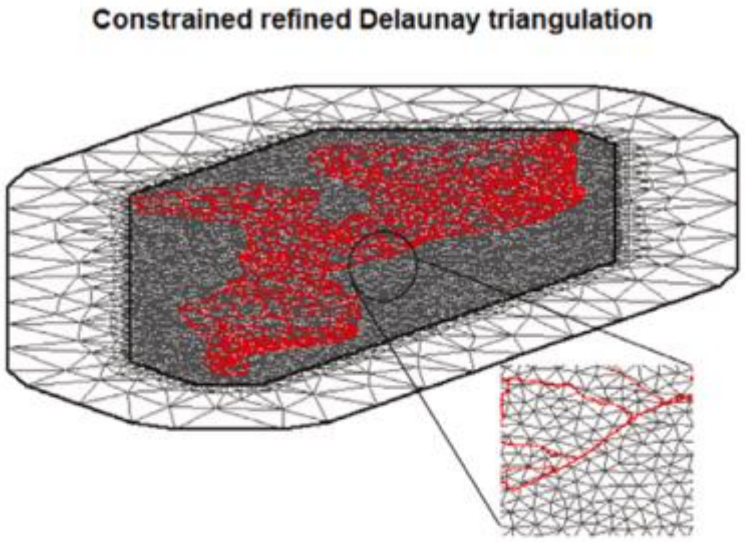
Why use Zenodo?

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display standards compliant usage statistics

Building the SPDE model

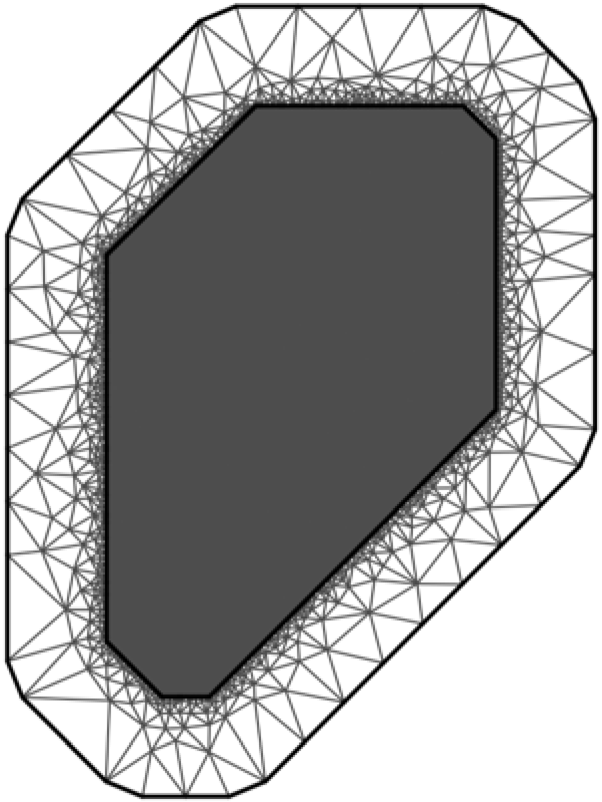


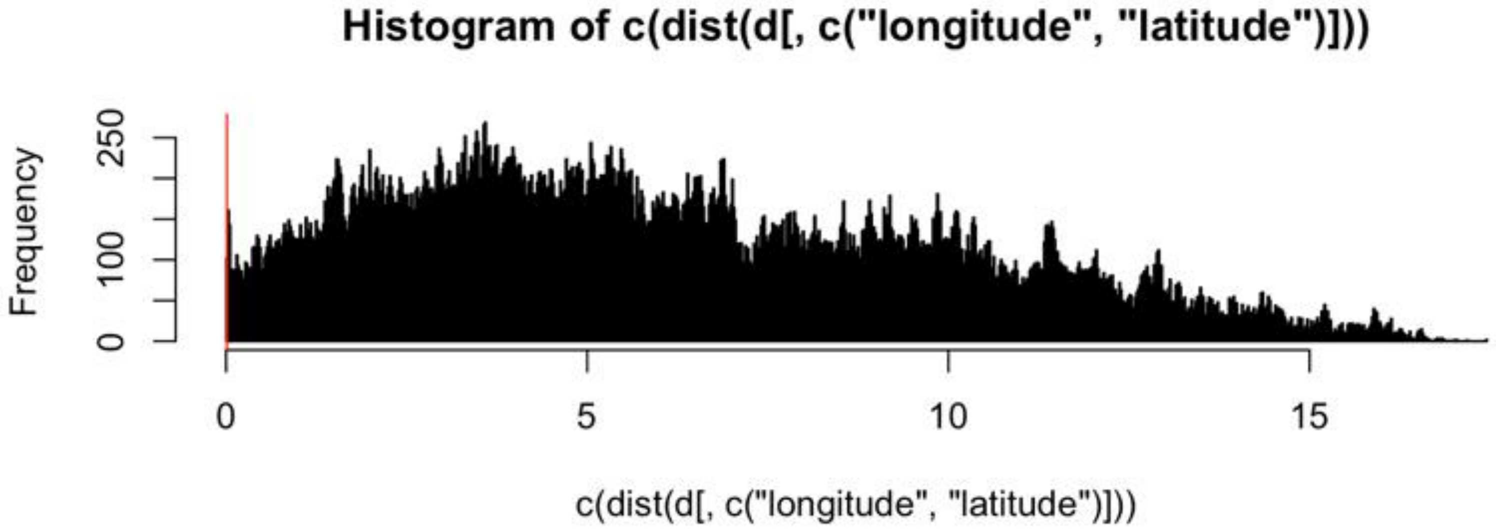
```
coo <- cbind(d$longitude, d$latitude)
mesh <- inla.mesh.2d(loc = coo, max.edge = c(0.1, 5),
      ↪  cutoff = 0.01)
```

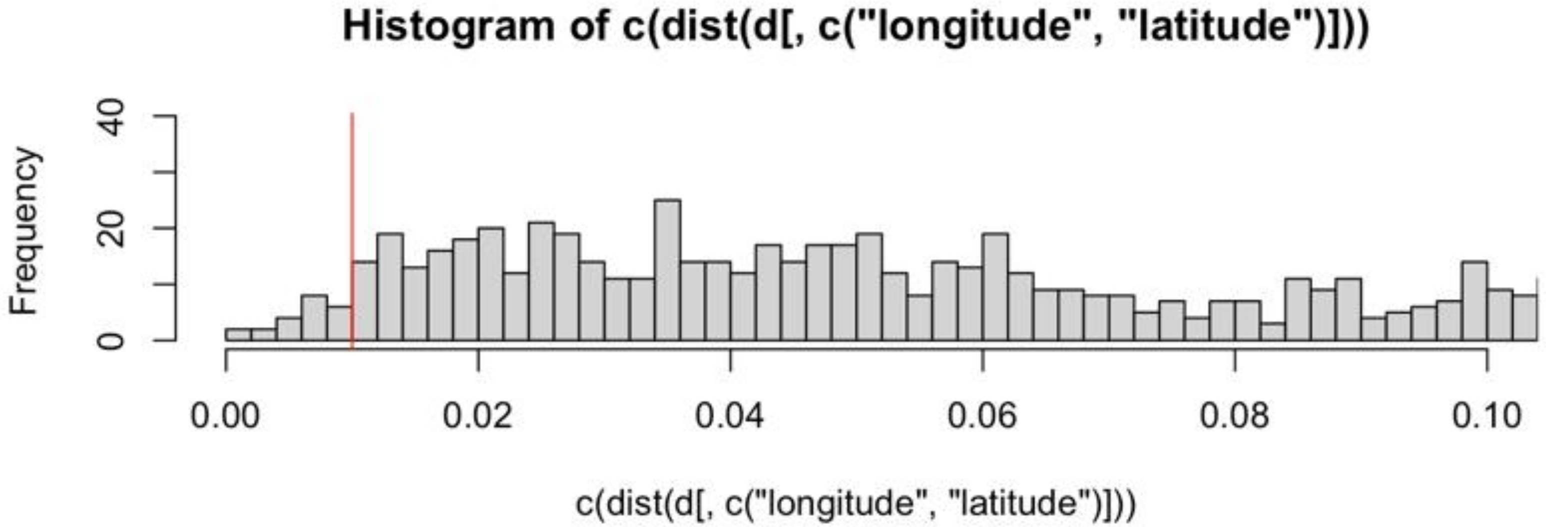


Constrained refined Delaunay triangulation

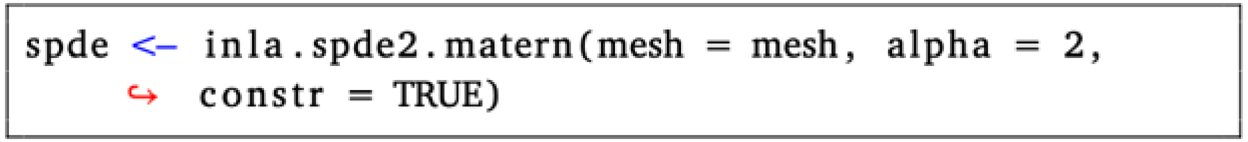
Fig. 2. Mozambique map overlaid on mesh used for model computation.



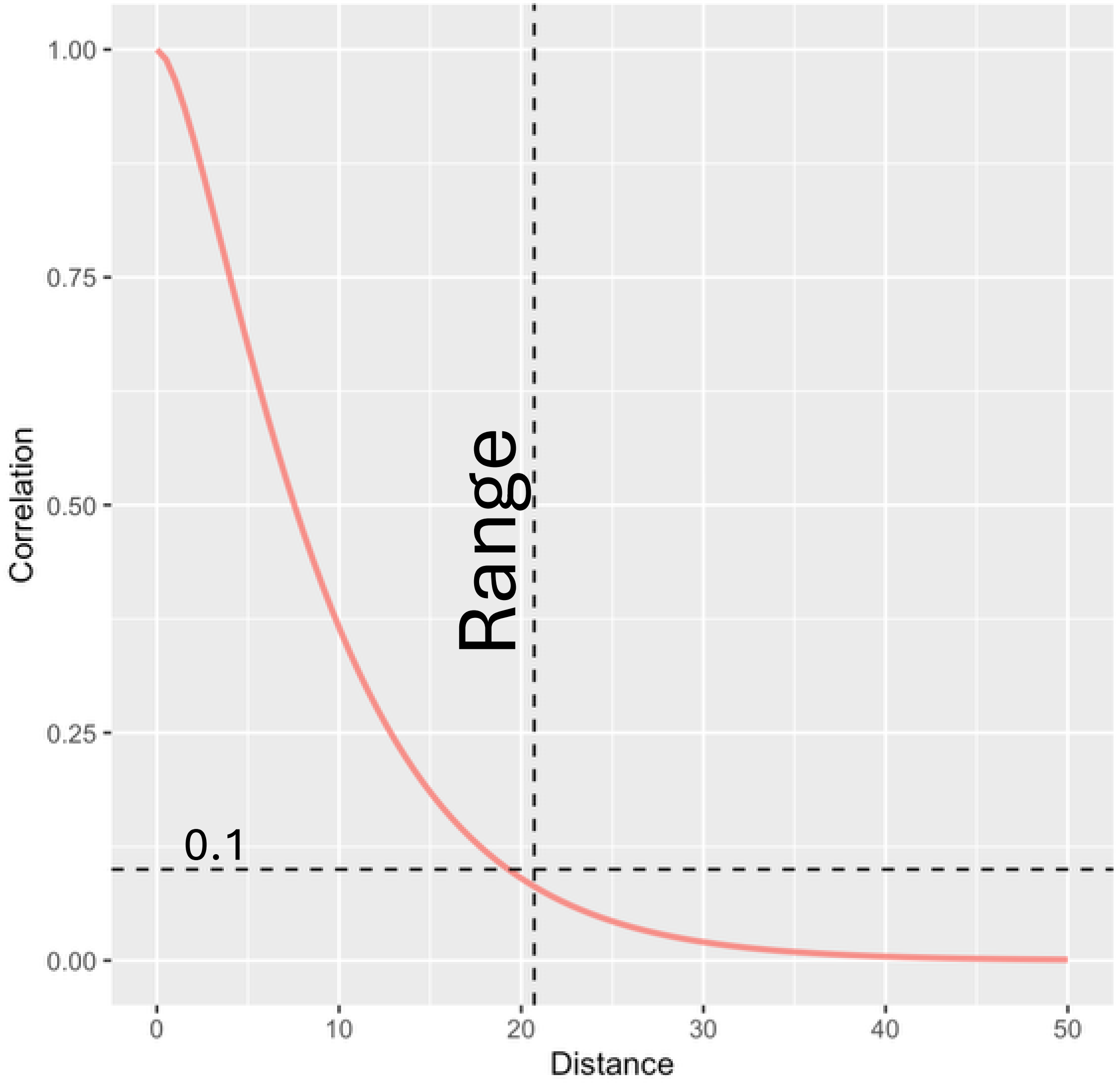

Histogram of c(dist(d[, c("longitude", "latitude")]))
Frequency
0
100
250
0
5
10
15
c(dist(d[, c("longitude", "latitude")]))


Histogram of c(dist(d[, c("longitude", "latitude")]))
Frequency
0
20
40
0.00
0.02
0.04
0.06
0.08
0.10
c(dist(d[, c("longitude", "latitude")]))

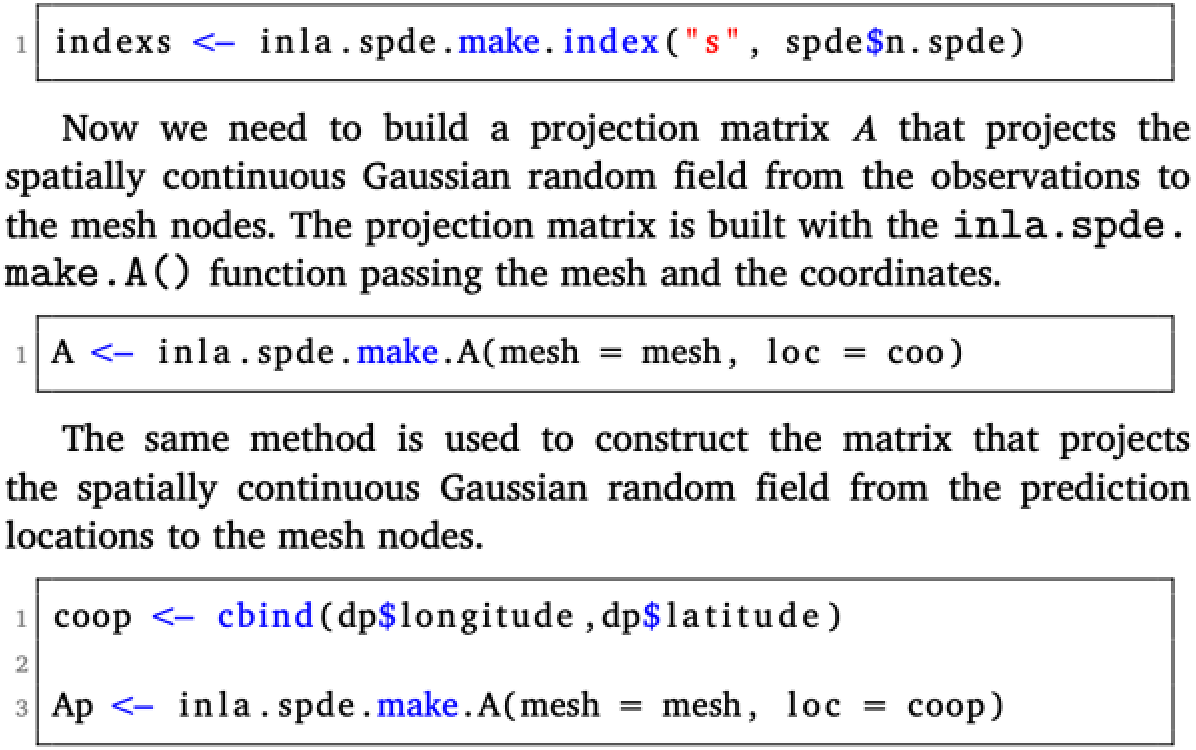
Building the SPDE model



```
spde <- inla.spde2.matern(mesh = mesh, alpha = 2,
     ↪  constr = TRUE)
```



Linking the SPDE model to the observations



```r
indexs <- inla.spde.make.index("s", spde$n.spde)
```

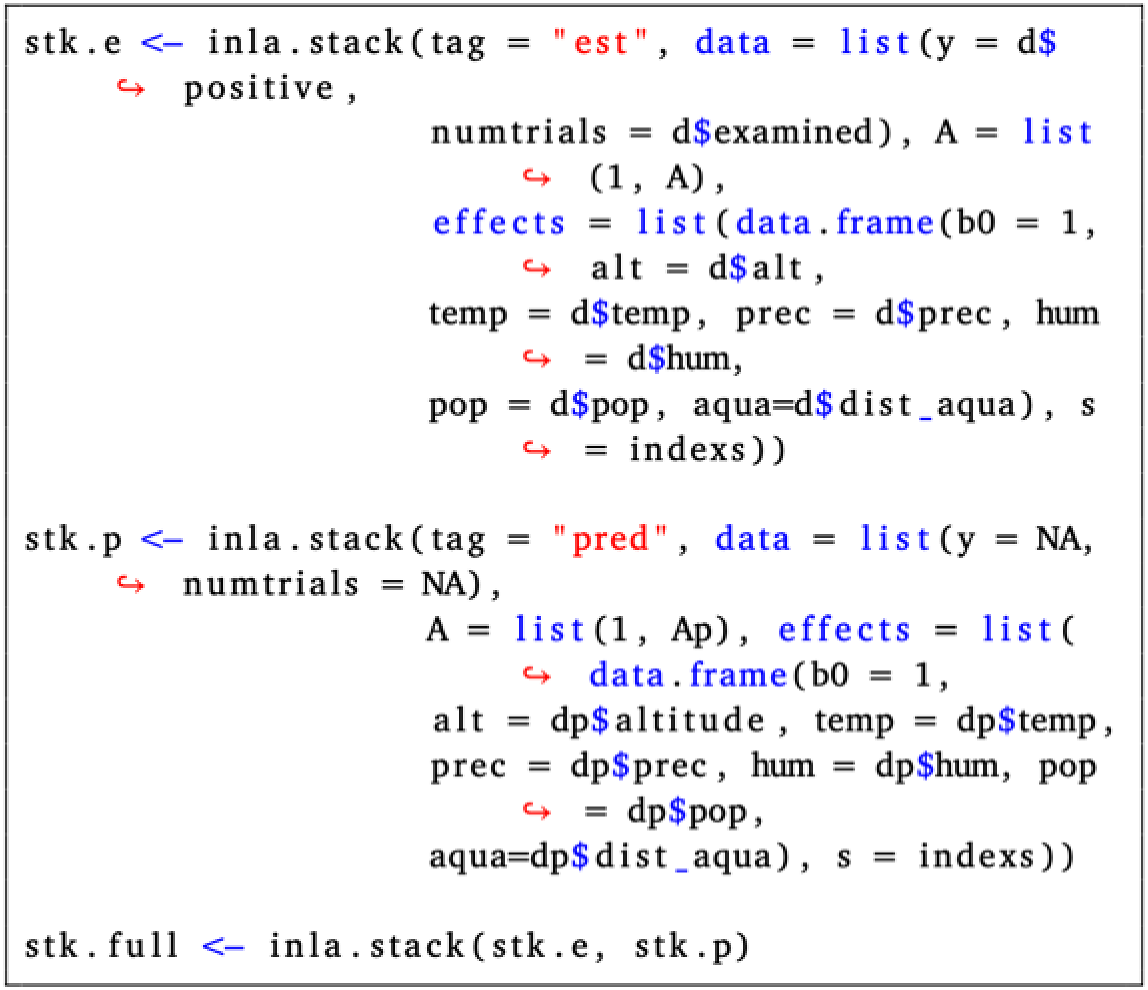
Now we need to build a projection matrix A that projects the spatially continuous Gaussian random field from the observations to the mesh nodes. The projection matrix is built with the `inla.spde.make.A()` function passing the mesh and the coordinates.

```r
A <- inla.spde.make.A(mesh = mesh, loc = coo)
```

The same method is used to construct the matrix that projects the spatially continuous Gaussian random field from the prediction locations to the mesh nodes.

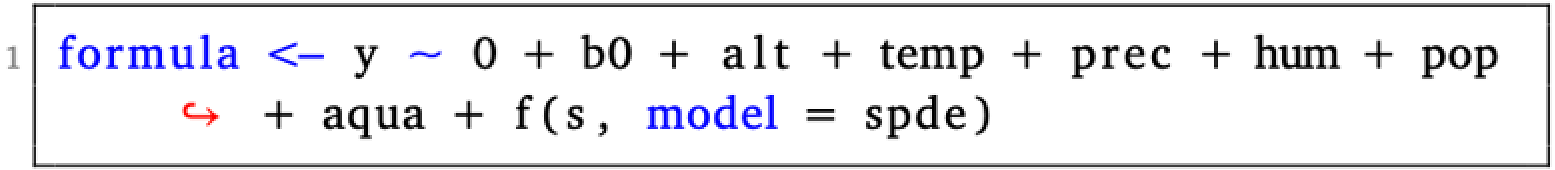
```r
coop <- cbind(dp$longitude,dp$latitude)

Ap <- inla.spde.make.A(mesh = mesh, loc = coop)
```

Building the stack (supplying all this data and structural info to INLA)

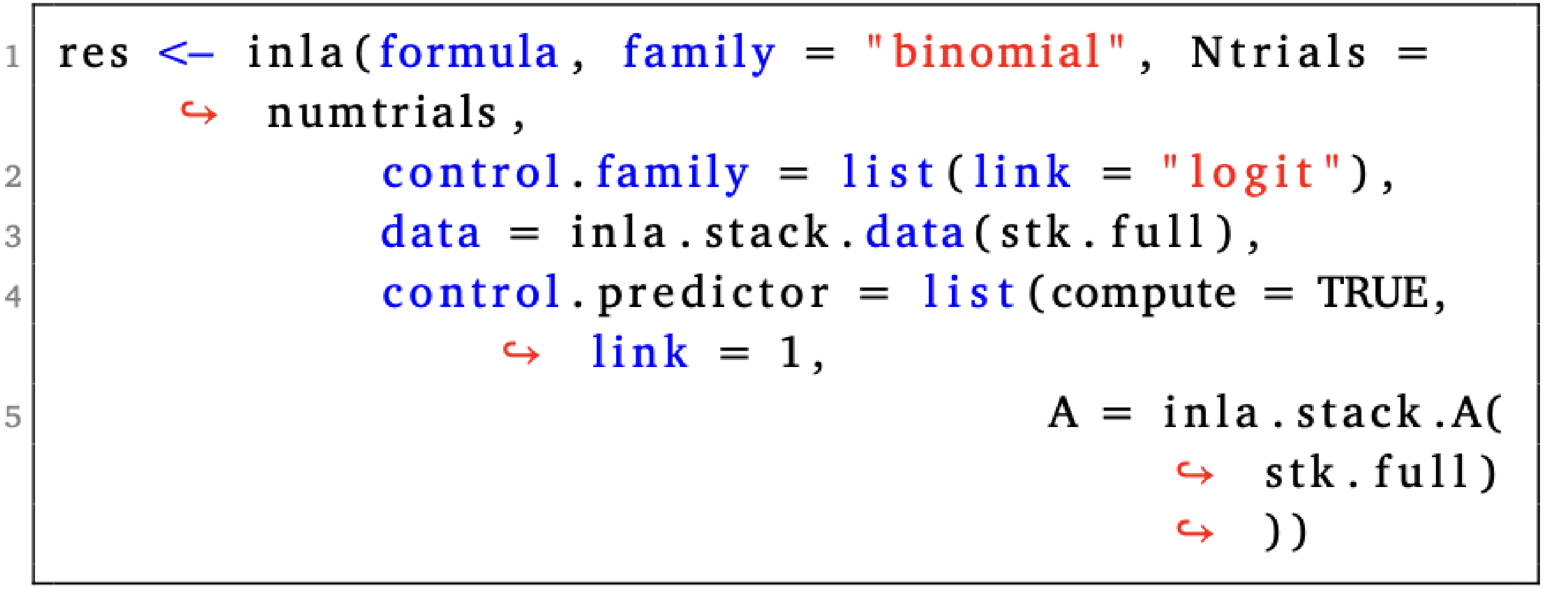


```r
stk.e <- inla.stack(tag = "est", data = list(y = d$positive,
                    numtrials = d$examined), A = list(1, A),
                    effects = list(data.frame(b0 = 1, alt = d$alt,
                    temp = d$temp, prec = d$prec, hum = d$hum,
                    pop = d$pop, aqua=d$dist_aqua), s = indexs))

stk.p <- inla.stack(tag = "pred", data = list(y = NA, numtrials = NA),
                    A = list(1, Ap), effects = list(data.frame(b0 = 1,
                    alt = dp$altitude, temp = dp$temp,
                    prec = dp$prec, hum = dp$hum, pop = dp$pop,
                    aqua=dp$dist_aqua), s = indexs))

stk.full <- inla.stack(stk.e, stk.p)
```

Finally specifying the formula and fitting the model



```r
formula <- y ~ 0 + b0 + alt + temp + prec + hum + pop + aqua + f(s, model = spde)
```



```r
res <- inla(formula, family = "binomial", Ntrials = numtrials,
            control.family = list(link = "logit"),
            data = inla.stack.data(stk.full),
            control.predictor = list(compute = TRUE, link = 1,
                                     A = inla.stack.A(stk.full)))
```

Results

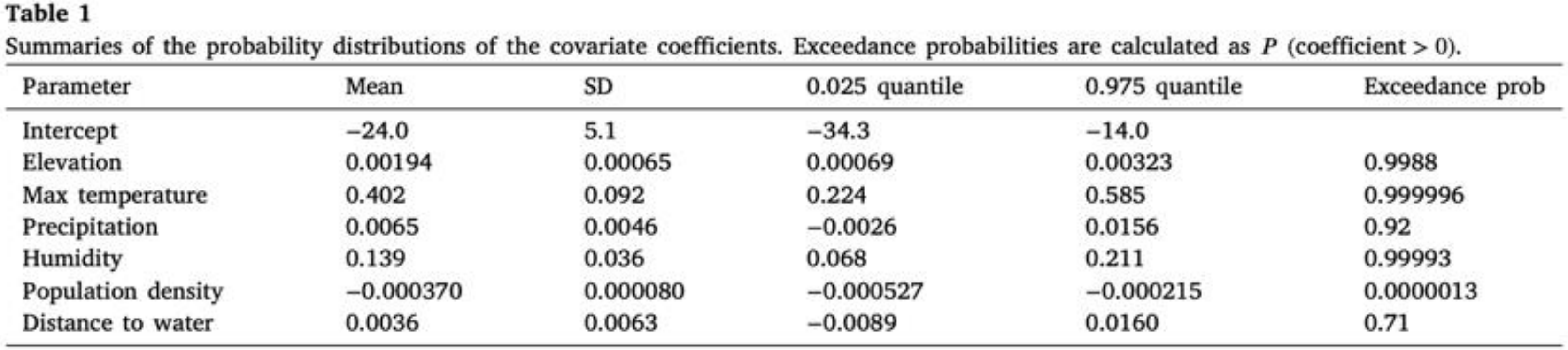


Table 1

Summaries of the probability distributions of the covariate coefficients. Exceedance probabilities are calculated as P (coefficient > 0).

Parameter	Mean	SD	0.025 quantile	0.975 quantile	Exceedance prob
Intercept	−24.0	5.1	−34.3	−14.0	
Elevation	0.00194	0.00065	0.00069	0.00323	0.9988
Max temperature	0.402	0.092	0.224	0.585	0.999996
Precipitation	0.0065	0.0046	−0.0026	0.0156	0.92
Humidity	0.139	0.036	0.068	0.211	0.99993
Population density	−0.000370	0.000080	−0.000527	−0.000215	0.0000013
Distance to water	0.0036	0.0063	−0.0089	0.0160	0.71

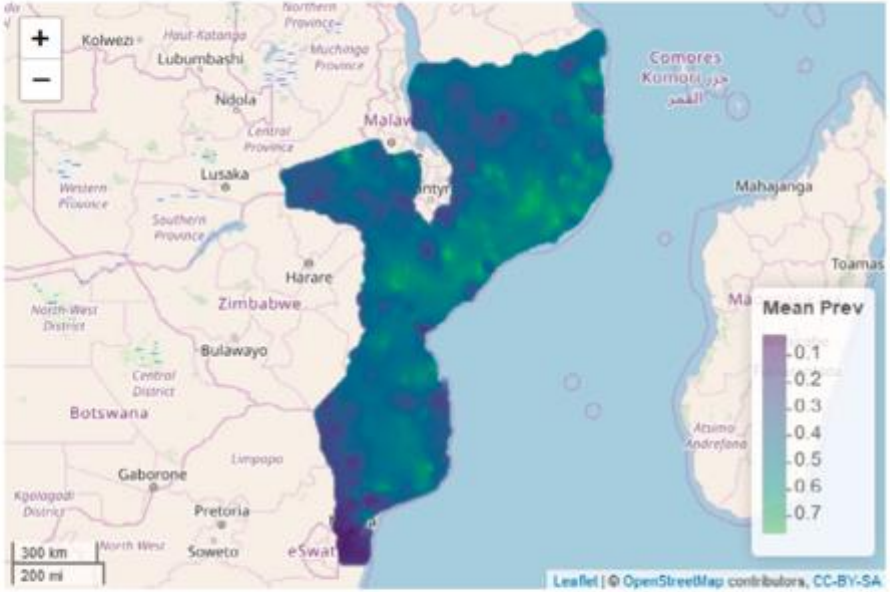
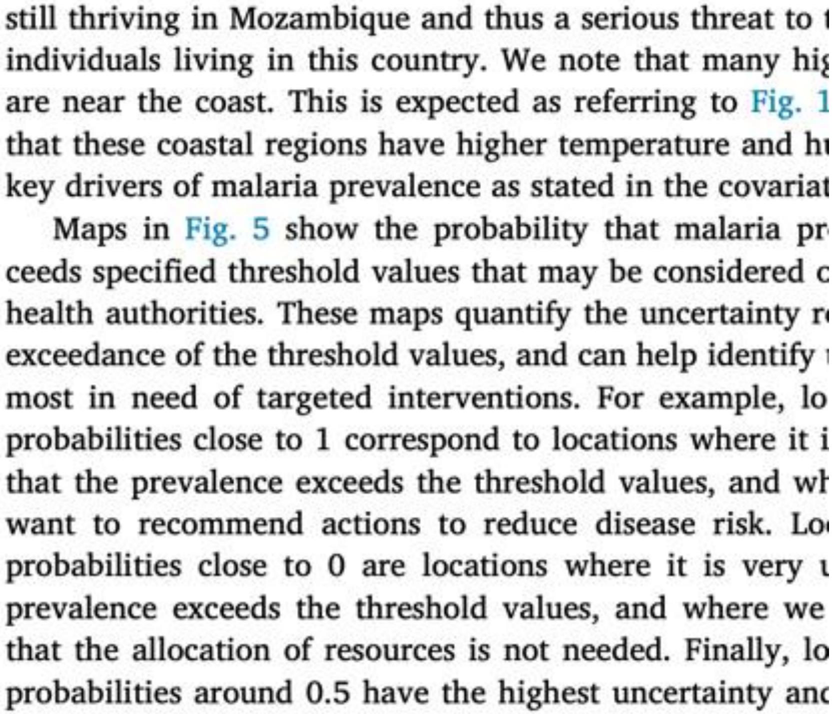


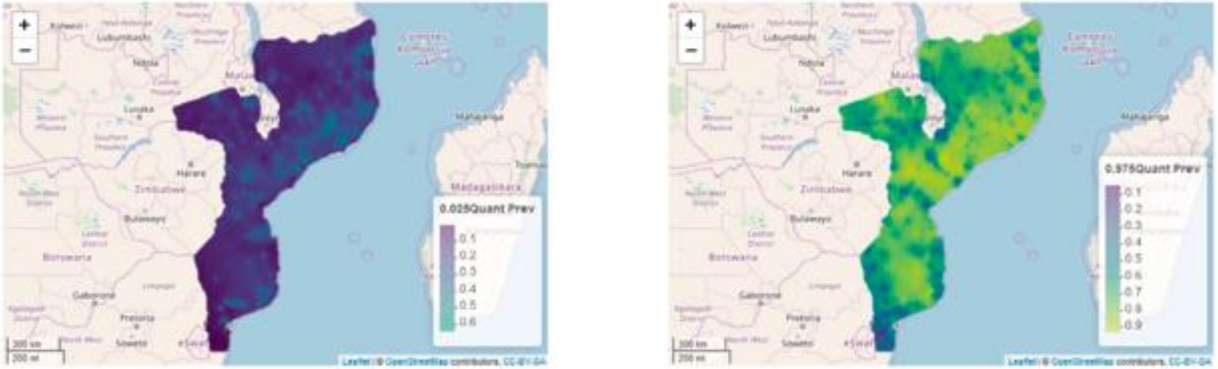
Fig. 3. Predicted mean values of prevalence of malaria across Mozambique for the fitted model.



still thriving in Mozambique and thus a serious threat to t individuals living in this country. We note that many hig are near the coast. This is expected as referring to Fig. 1 that these coastal regions have higher temperature and hu key drivers of malaria prevalence as stated in the covariat

Maps in Fig. 5 show the probability that malaria pr ceeds specified threshold values that may be considered c health authorities. These maps quantify the uncertainty r exceedance of the threshold values, and can help identify t most in need of targeted interventions. For example, lo probabilities close to 1 correspond to locations where it i that the prevalence exceeds the threshold values, and wh want to recommend actions to reduce disease risk. Lo probabilities close to 0 are locations where it is very u prevalence exceeds the threshold values, and where we that the allocation of resources is not needed. Finally, lo probabilities around 0.5 have the highest uncertainty anc

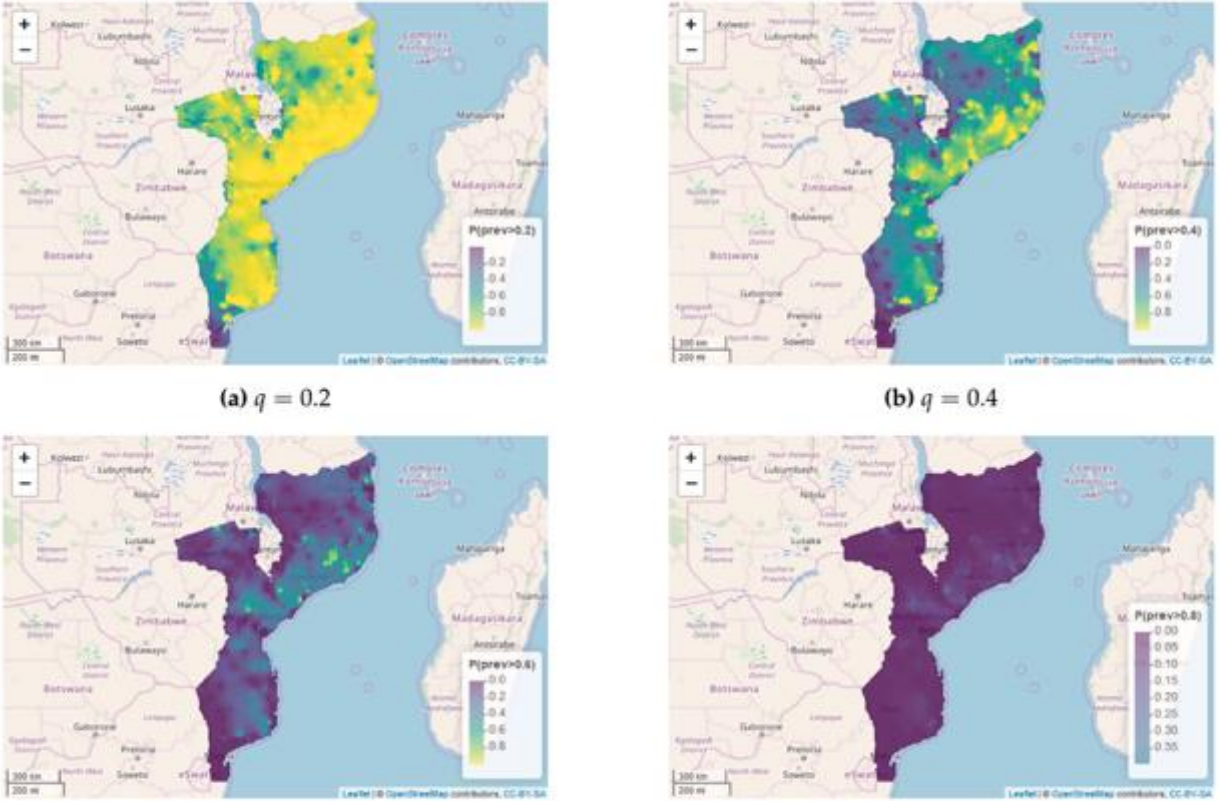
Results



(a) Predicted lower limit (0.025 quantile) values of prevalence of malaria across Mozambique for our fitted model.

(b) Predicted upper limit (0.975 quantile) values of prevalence of malaria across Mozambique for our fitted model.

Fig. 4. Upper and lower limits of prevalence.



(a) $q = 0.2$

(b) $q = 0.4$

(c) $q = 0.6$

(d) $q = 0.8$

Results

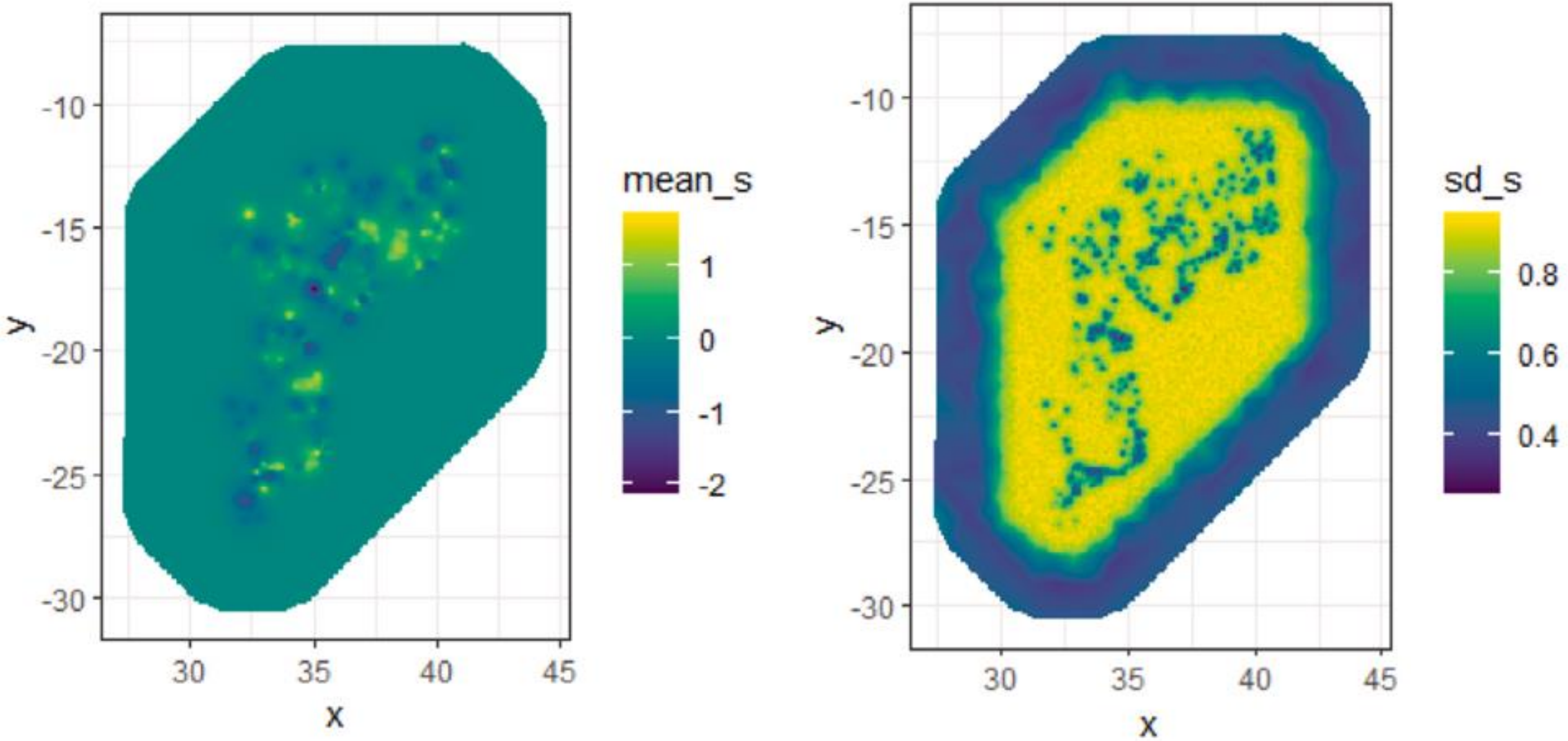
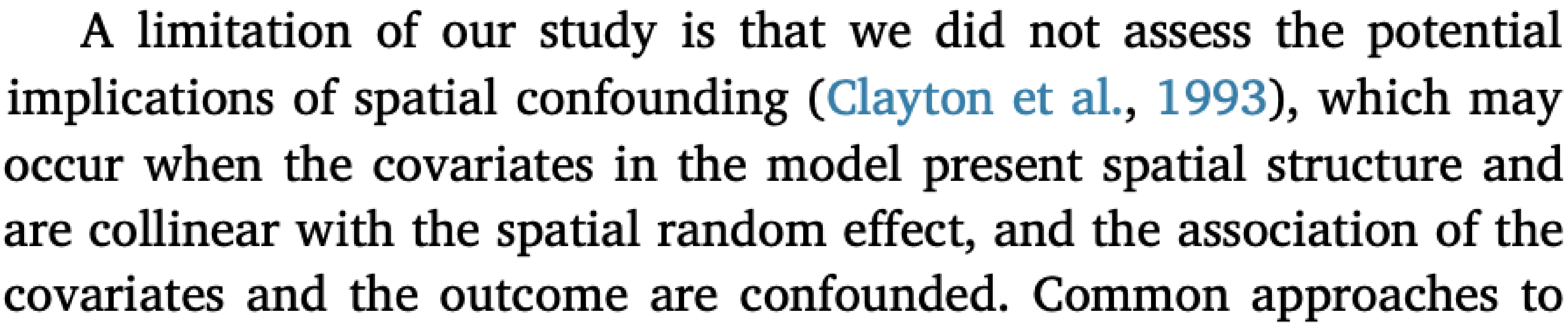
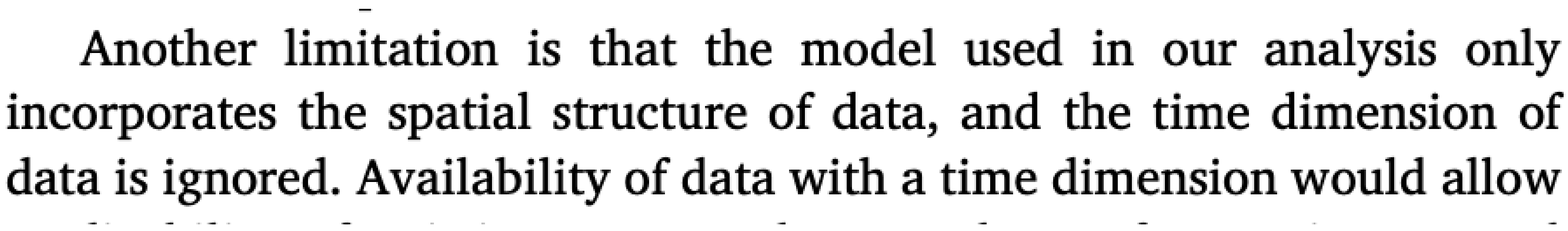


Fig. 6. Mean and SD of the spatial field across our mesh.

Discussion



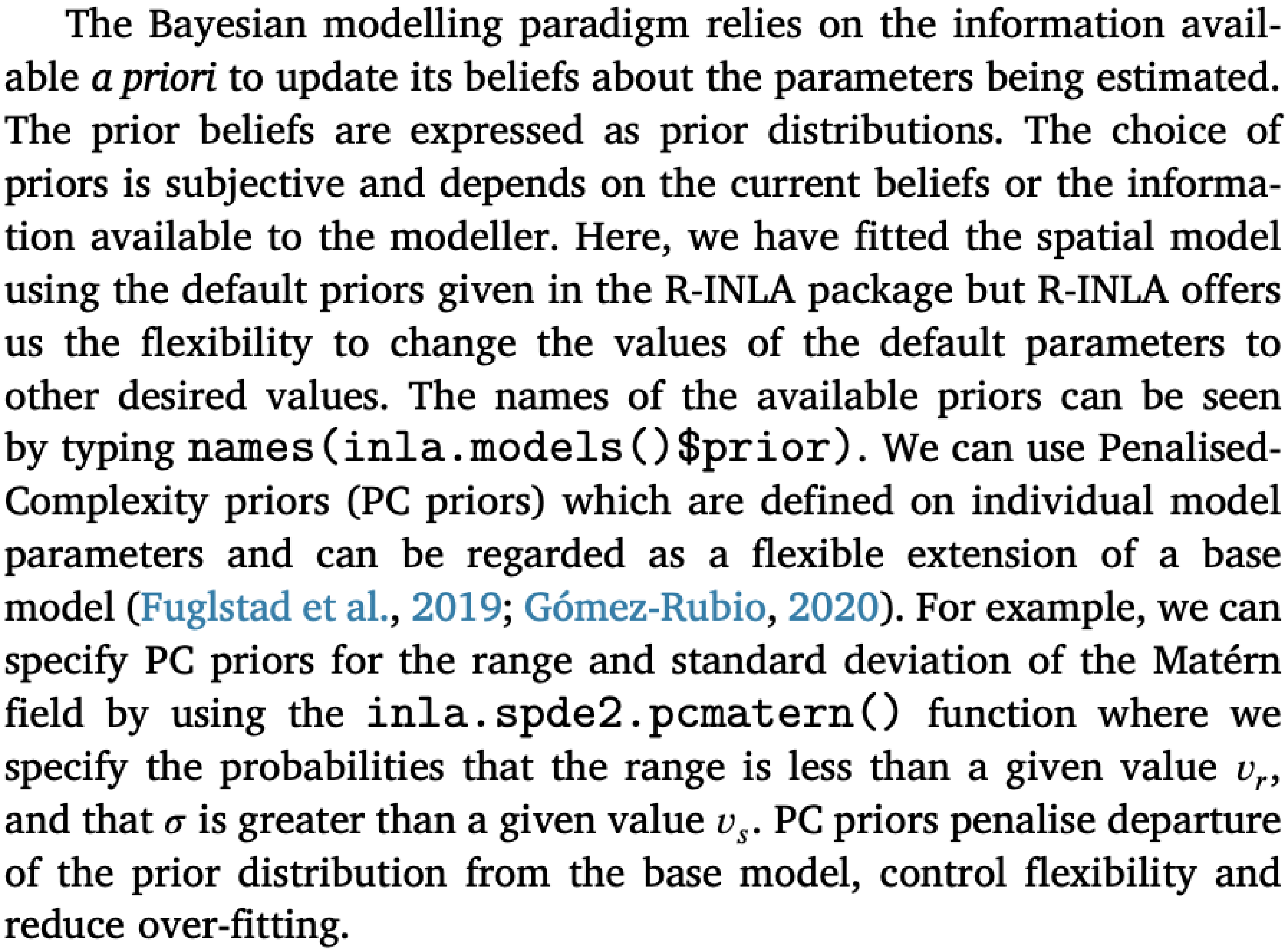
A limitation of our study is that we did not assess the potential implications of spatial confounding (Clayton et al., 1993), which may occur when the covariates in the model present spatial structure and are collinear with the spatial random effect, and the association of the covariates and the outcome are confounded. Common approaches to



Another limitation is that the model used in our analysis only incorporates the spatial structure of data, and the time dimension of data is ignored. Availability of data with a time dimension would allow

Discussion

- Priors…



The Bayesian modelling paradigm relies on the information available *a priori* to update its beliefs about the parameters being estimated. The prior beliefs are expressed as prior distributions. The choice of priors is subjective and depends on the current beliefs or the information available to the modeller. Here, we have fitted the spatial model using the default priors given in the R-INLA package but R-INLA offers us the flexibility to change the values of the default parameters to other desired values. The names of the available priors can be seen by typing `names(inla.models()$prior)`. We can use Penalised-Complexity priors (PC priors) which are defined on individual model parameters and can be regarded as a flexible extension of a base model (Fuglstad et al., 2019; Gómez-Rubio, 2020). For example, we can specify PC priors for the range and standard deviation of the Matérn field by using the `inla.spde2.pcmatern()` function where we specify the probabilities that the range is less than a given value v_r, and that σ is greater than a given value v_s. PC priors penalise departure of the prior distribution from the base model, control flexibility and reduce over-fitting.

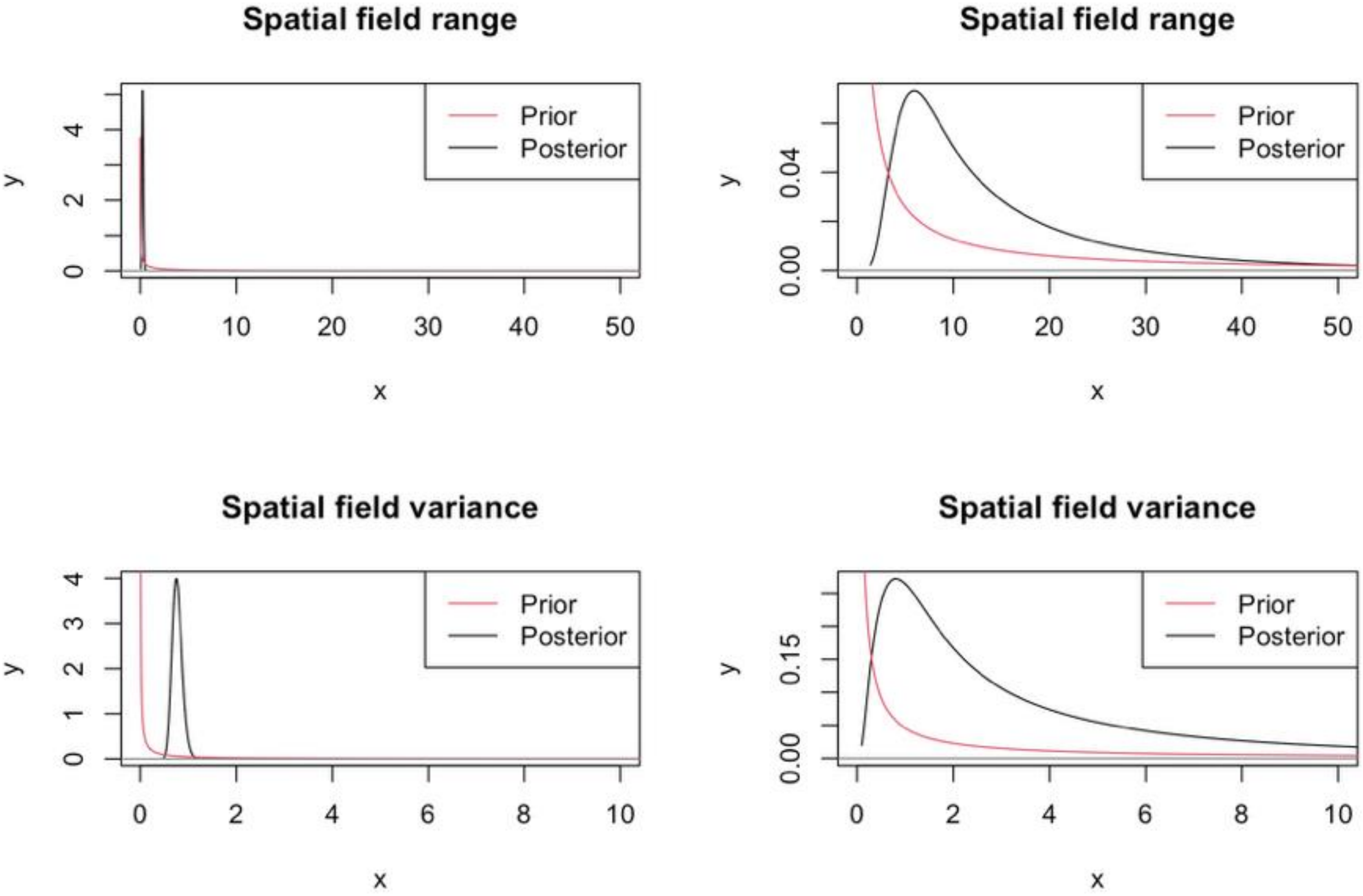
A closer look at the spatial random effect results

- What is missing from this model?

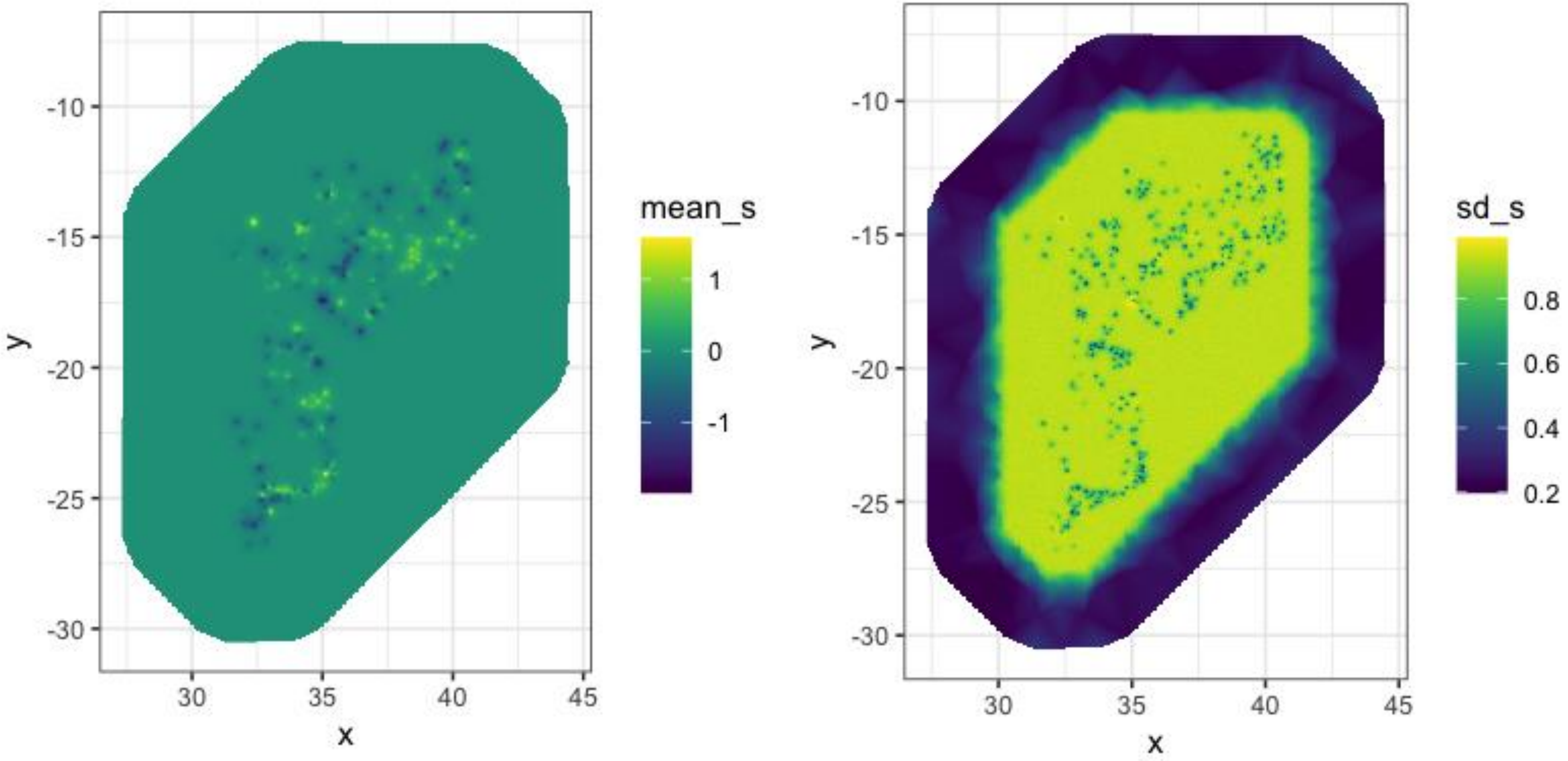
$$\operatorname{logit}\left(P\left(\mathbf{x}_i\right)\right) = \beta_0 + D_i\boldsymbol{\beta} + S(\mathbf{x}_i),$$

- A dispersion parameter

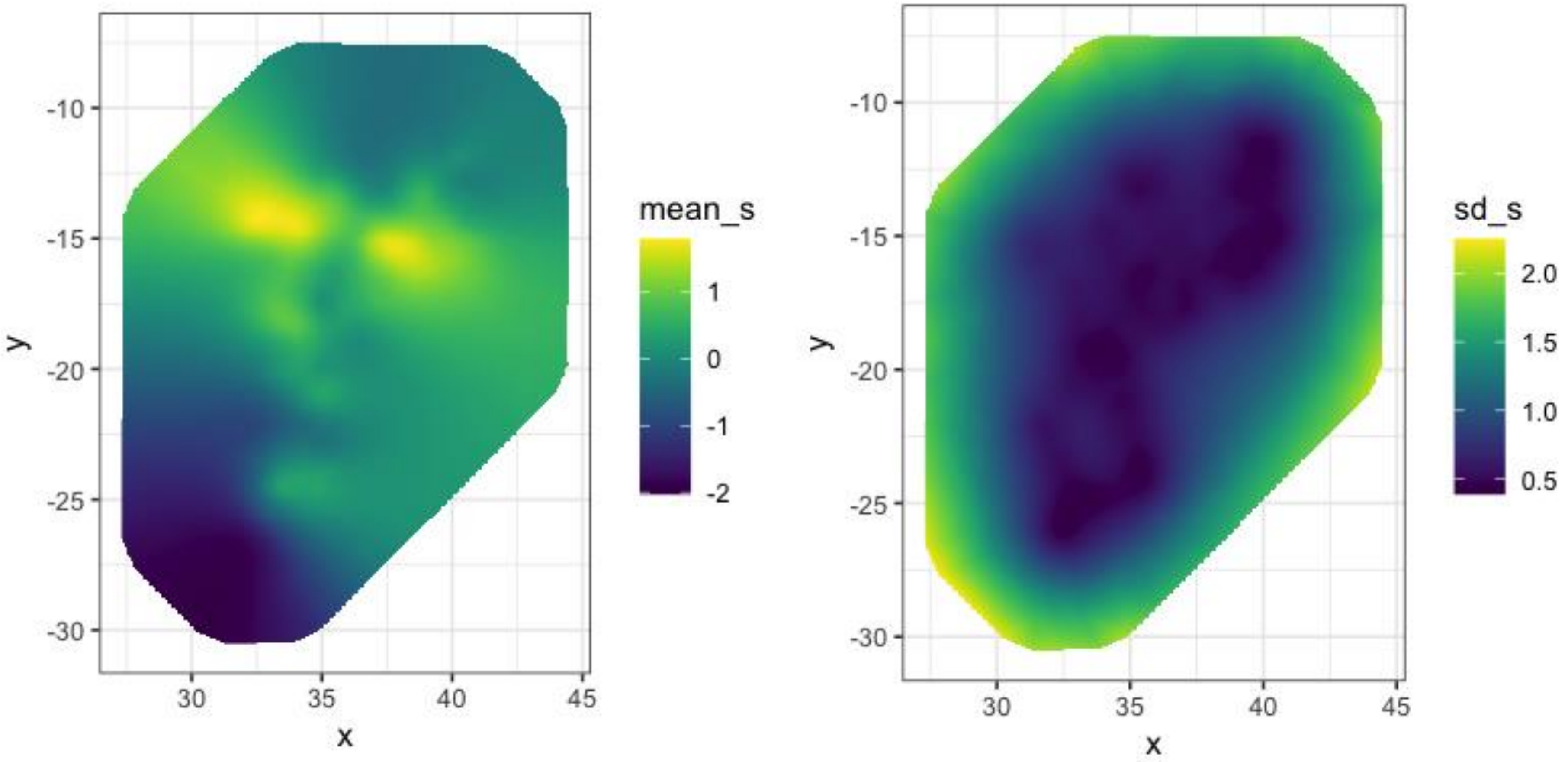
Adding a dispersion parameter (a row-level IID random effect) has a huge effect on the range of the spatial random effect. The mode of the range posterior changes from highly local ~0.1 degree (top left plot) to ~7 degrees (top right plot). Remember that the total range of the study area is only 10-15 degrees. There is also much more uncertainty about the range and variance. What I believe is happening is that in the absence of a dispersion parameter, the spatial field is being forced to model this substantial local variation. When the dispersion parameter is added, the spatial random effect is freed to model the “true” scale of spatial correlation.



Original model. Correlation is short range, so the spatial part of the model can only predict very close to where the data is, as shown by the high uncertainty (yellow, right plot) everywhere except right next to sampled locations.

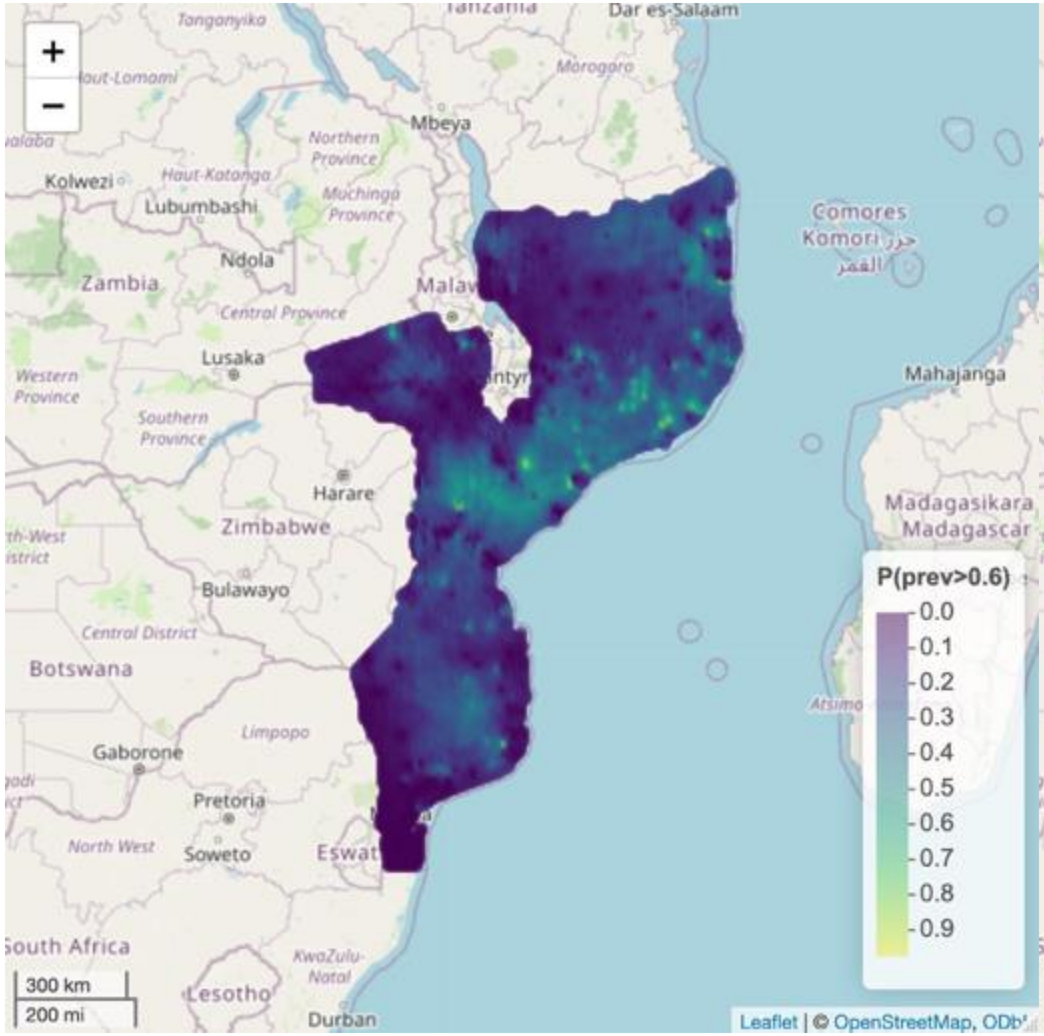


Original model + dispersion parameter. Correlation is long range, so the spatial part of the model is able to predict away from sampled locations with relatively high certainty.

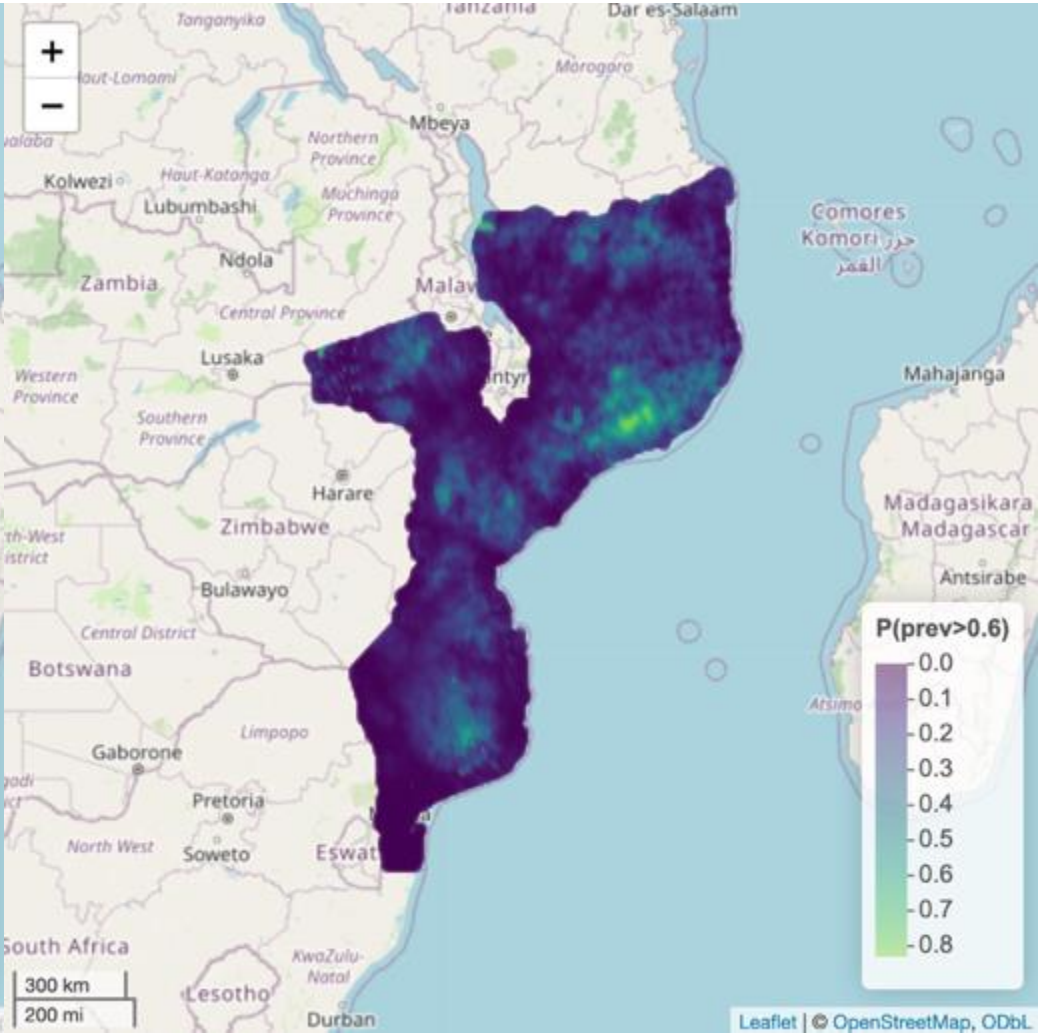


This has consequences for calculating exceedance probabilities, e.g. the areas at highest risk of exceeding 60% prevalence differ between the two models.

Original model



Original model + dispersion parameter



Note: These two models differ in more than just the presence of a dispersion parameter. I also logged the skewed covariates to reduce the influence of outliers and reduce linear correlations between the covariates, and centred and scaled them, which makes using the default priors on the intercept and other fixed effects more justifiable. Transforming the covariates reduced WAIC by ~13 points. Adding a dispersion parameter reduced WAIC by ~100 points.

See https://github.com/pcdjohnson/spatial-model-malaria for details.