Spatial Modelling of Netherlands precipitation

Spatial Modelling of Netherlands precipitation using three methods variogram, Gaussian Process model using maximum likelihood, and Bayesian model.

The dataset contains 220 measurements of total monthly precipitation in the Netherlands in September 2019. This dataset was downloaded from the Copernicus Climate Data Store [1]. In the dataset, each row contains a station name, longitude and latitude of the observation station, and total precipitation for the month in millimetres.

The work will identify the spatial relationships seen in the data, model the precipitation and the spatial variations and predict precipitation at 3 random locations. Three methods are used to predict: variogram, Gaussian Process model using maximum likelihood, and Bayesian model.

[1] Copernicus Climate Change Service, Climate Data Store, (2021): Global land surface atmo- spheric variables from 1755 to 2020 from comprehensive in-situ observations. Copernicus Cli- mate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.cf5f3bac (Accessed on 23-MAR-2023)

Exploratory Data Analysis

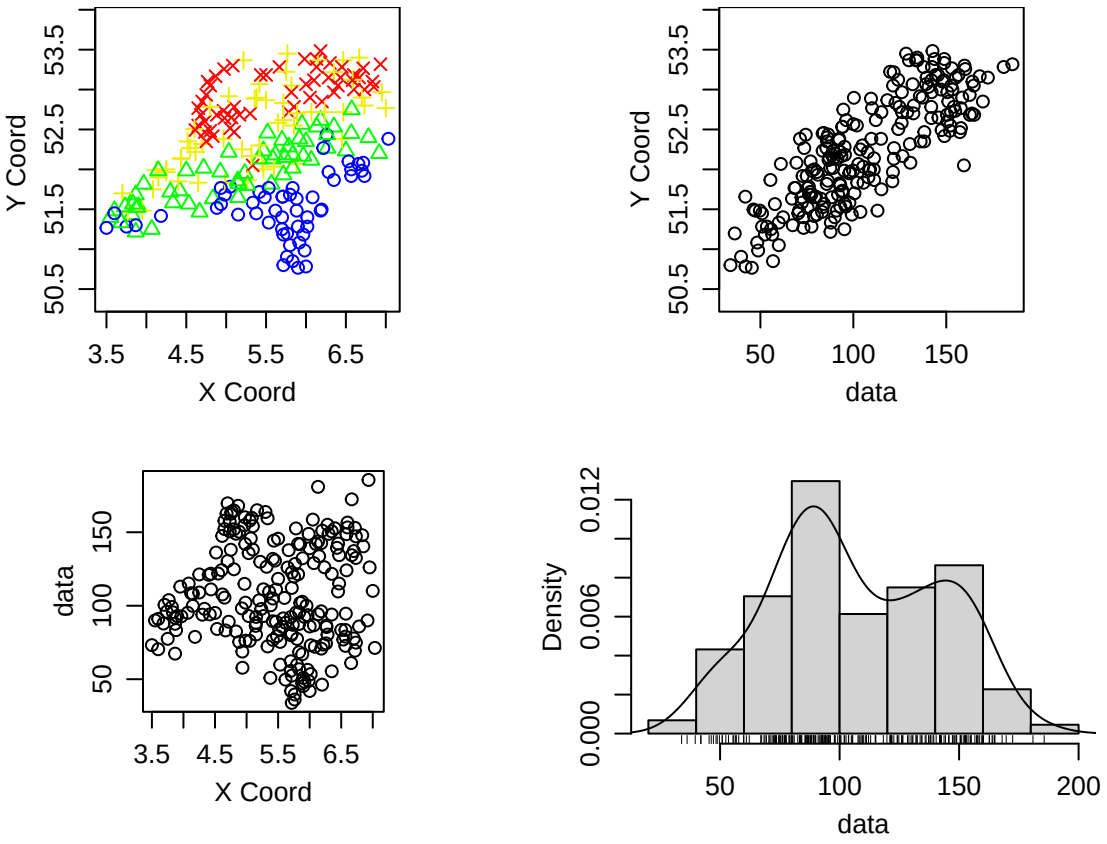


Figure 1: Plot the precipitation of Netherlands in September 2019

The top right plots the precipitation against latitude. It can be seen that higher latitude recorded higher precipitation. This reflects that northern Netherlands recorded higher precipitation than the southern.

The bottom left plots the precipitation against longitude. Rainfall volume seems to widely vary at right longitude, while at left-most longitude they center around 100. This can be due to the fact that there are fewer observations recorded there.

The top left plots the precipitation all over Netherlands, with blue, green, yellow and red respectively represent the lowest, second, third and highest quartile in the precipitation range. As mentioned, the southern region

is covered in blue, while moving up to the north, it gradually changes to green, yellow and then red in the north of Netherlands.

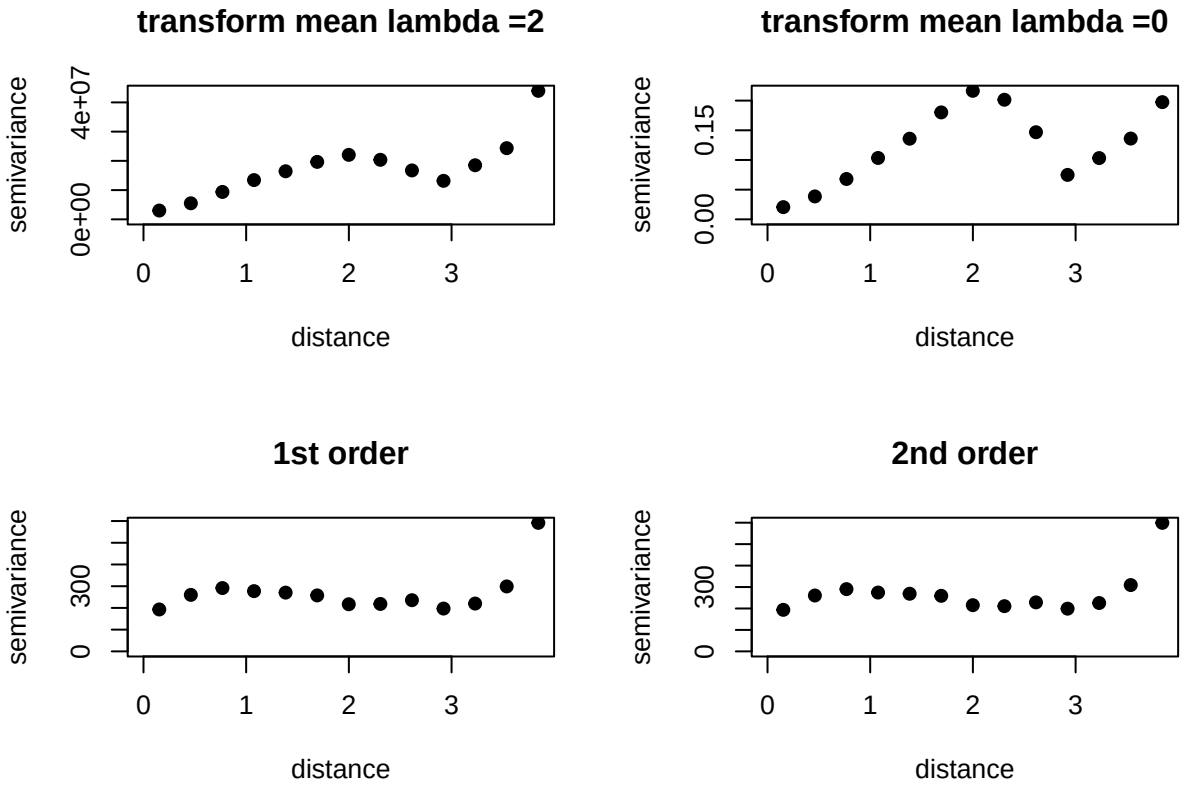
There is clearly spatial correlation. Regions of the same latitude seem to have similar or nearly similar precipitation, while precipitation of regions of the same longitude can vary (shown in the mixture of blue, green and yellows points in the same longitude)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
33.9	80.85	100.15	106.5705	137.35	185.6

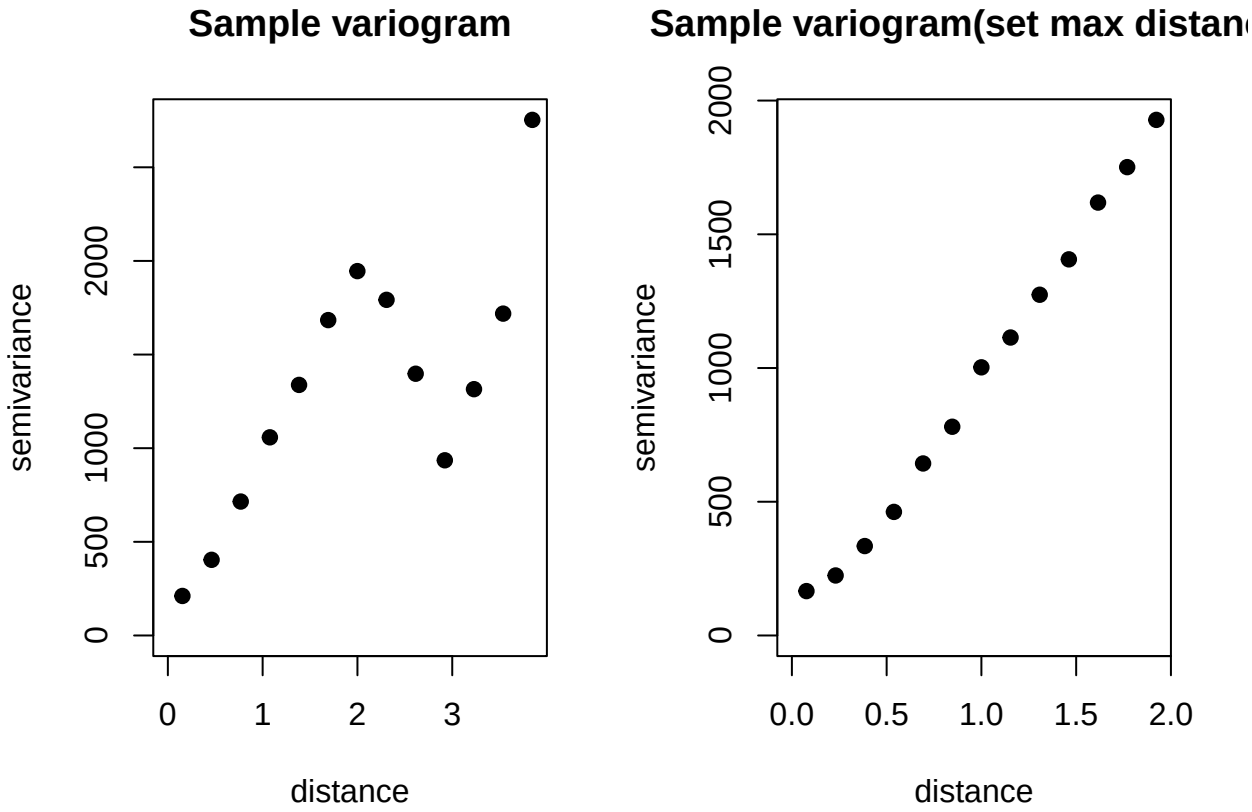
Figure 2: Numerical summary

From the numerical summary and density plot, precipitation ranges from nearly 34 to over 185, with mean = 100 and median = 106.

Variogram



The bottom variograms when we assume the mean has a first or second order polynomial on the coordinates are not stable and continuous. The variance fluctuates over the range of distances and does not reach a sensible sill. The top variograms when we transform the mean share the same trend with the original variogram (with assumption of constant mean by default).



We use the original variogram. We need to set a maximum distance = 2 before fitting a model because when the distance is larger than 2, the variance starts to drop. This decrease does not describe the variance of points which are far apart, but due to the fact that there are not as many data points of this distance as of closer distance. Thus, from distance = 2, variance starts to drop before going up at 3.

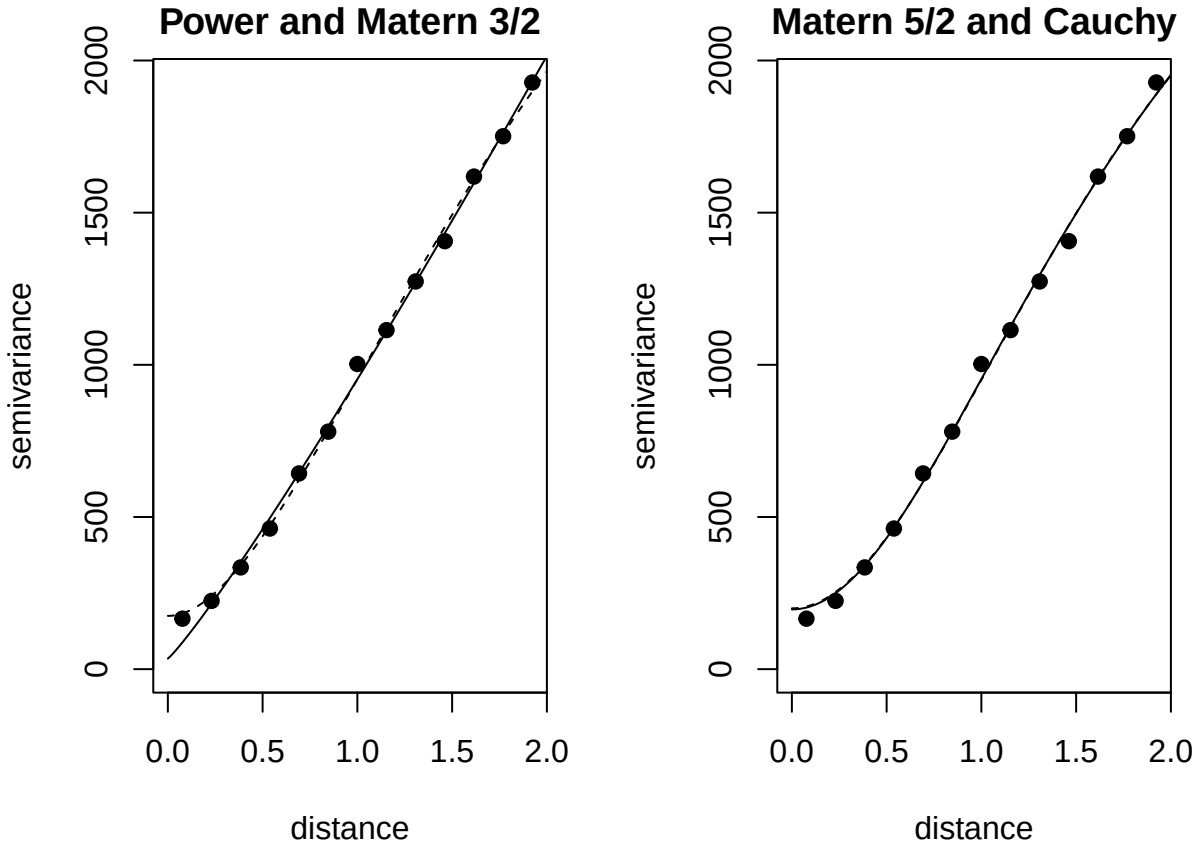
We need to include a nugget. As can be seen from the right plot, the variance at distance = 0 is larger than 0 (could be around 100), therefore a nugget is needed to reflect this variance.

	model	min_sum_sq
7	power	11181319
11	mattern 3/2	11853782
9	cauchy	14806997
12	mattern 5/2	15767716
5	cubic	18264745
10	gneiting	22376235
2	gaussian	23263283
4	circular	24095471
3	spherical	24095602
1	exponential	24174761
6	wave	28534445
13	mattern est	30819663
8	powered.exponential	1061231062

Figure 3: Covariance functions result

By this we assume that the mean function is constant (variogram by default). In the earlier part we have tested with other assumptions of mean, but ending up with constant mean assumption over the region.

We iterate the variogram with different assumptions about covariance functions, with fitted nugget and let it estimate the paramteters using weighted least squares. With Matern model, we test some assumptions of smoothness parameter kappa, including 3/2 and 5/2. With other models, we do not force the nugget to be zero, and fit the variogram by assuming different covariance models. Comparing different fitted models by the minimised weighted sum of squares, we can see that the Covariance Model = Power, Matern 3/2, Matern 5/2 and Cauchy result in the least residuals. We plot these fitted models to our sample variogram.

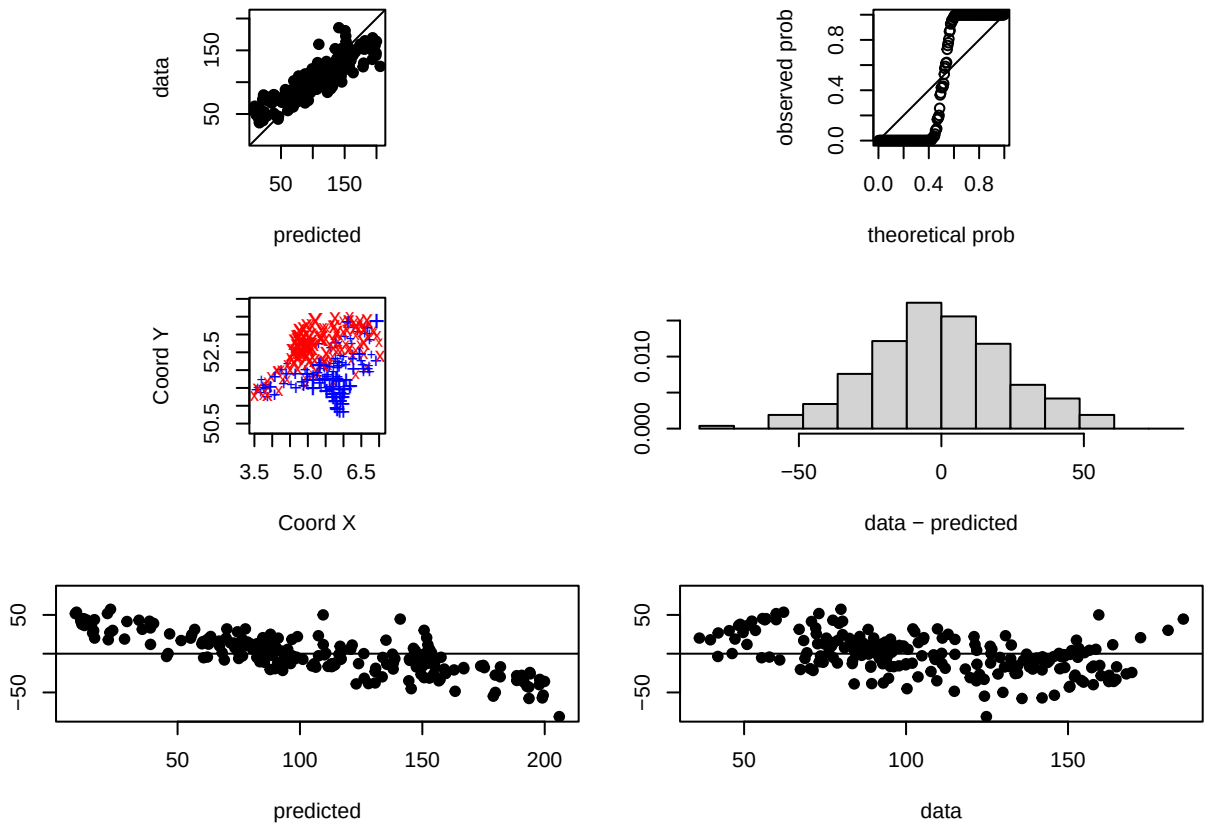


The matern 3/2 captures better the variance at close distance, while the power underestimate the nugget but looks better at far distance (when distance >1). The Matern 5/2 and Cauchy are quite identical, and both slightly overestimate the nugget. Except for the Power which estimates the nugget at around 34, The matern 3/2 estimates the nugget at over 174 and the other two models estimate the nugget at more than 190. The correlation length of Power and Matern 3/2 models are quite similar, 1.1 (Power) and Matern 3/2 (1.05). Meanwhile, Cauchy returns high correlation length (1.4) and Matern 5/2 returns much lower correlation length (0.6).

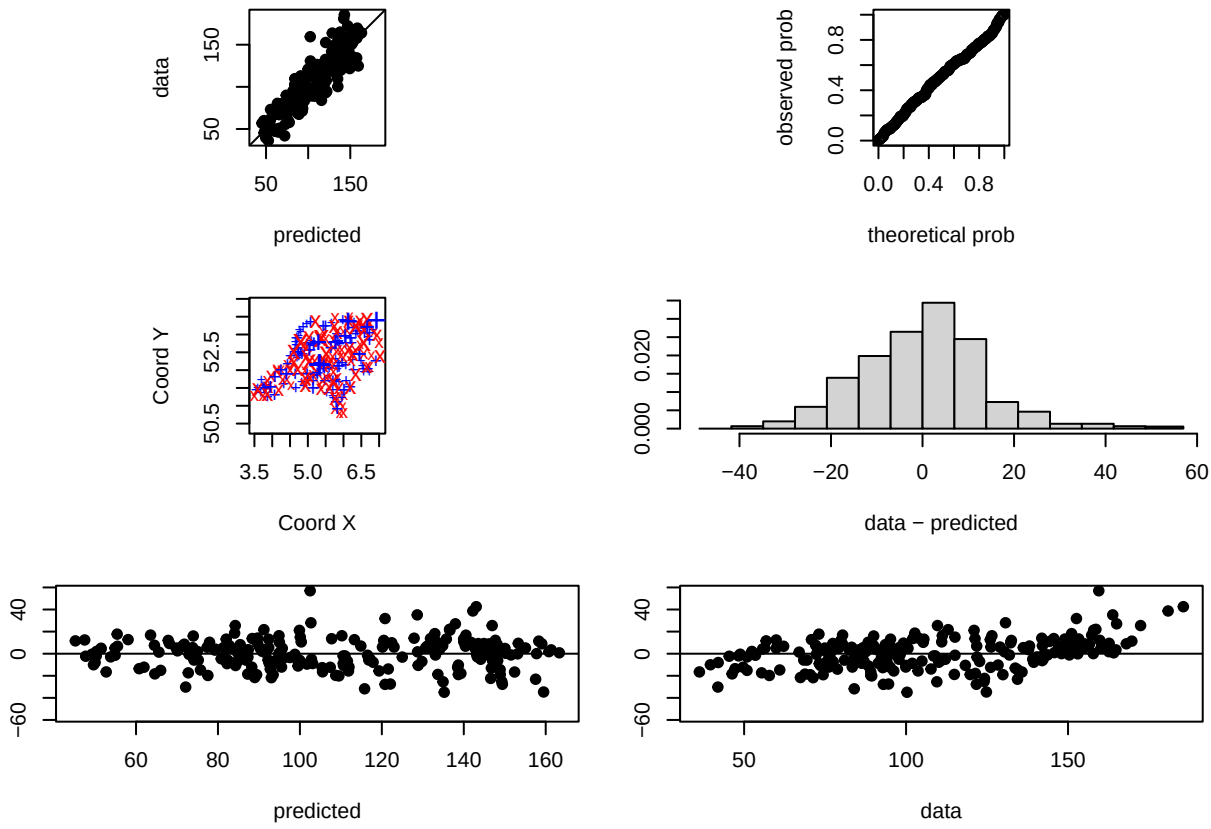
Out of the 4 models, Power does not have a sensible partial sill (918, much lower than 2000). The other three models all have the partial sill larger than 2500, which resonates in the variogram plot.

Validate variogram model: Power, Matern 3/2, Matern 5/2 and Cauchy

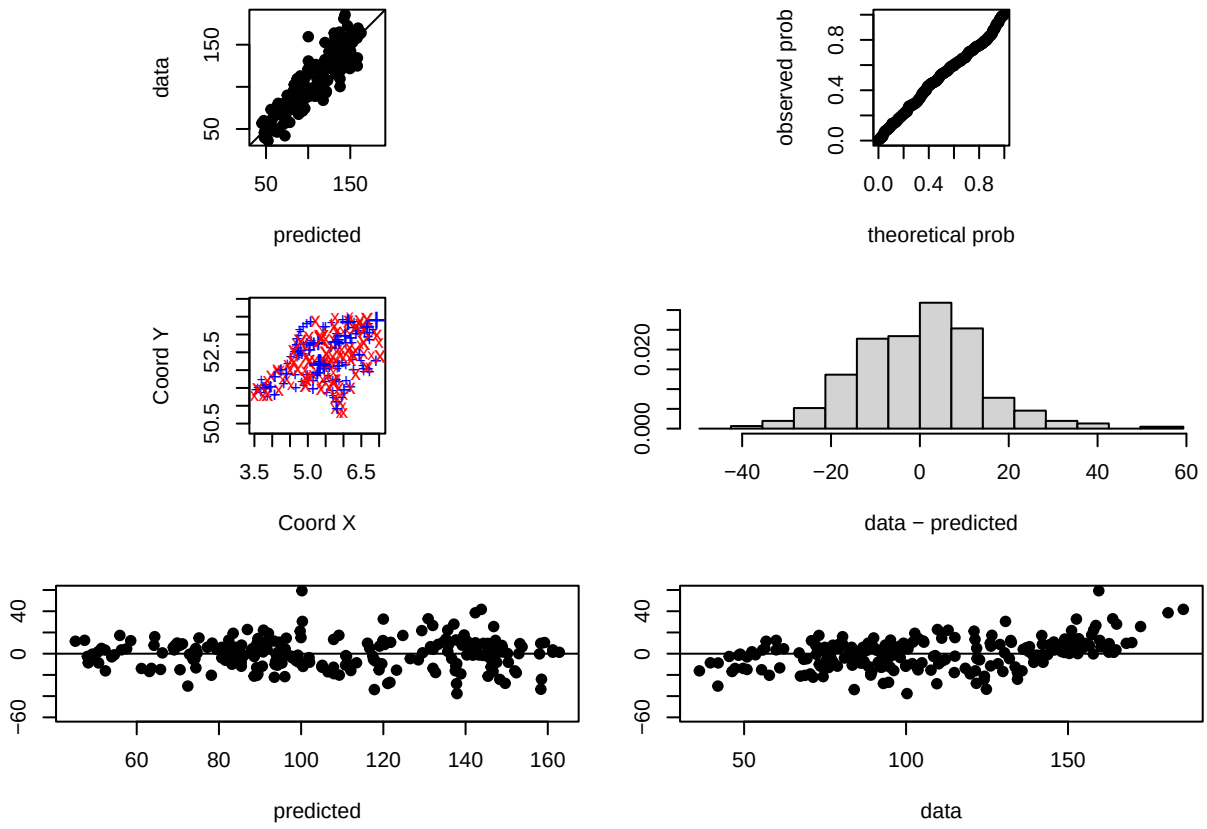
```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```



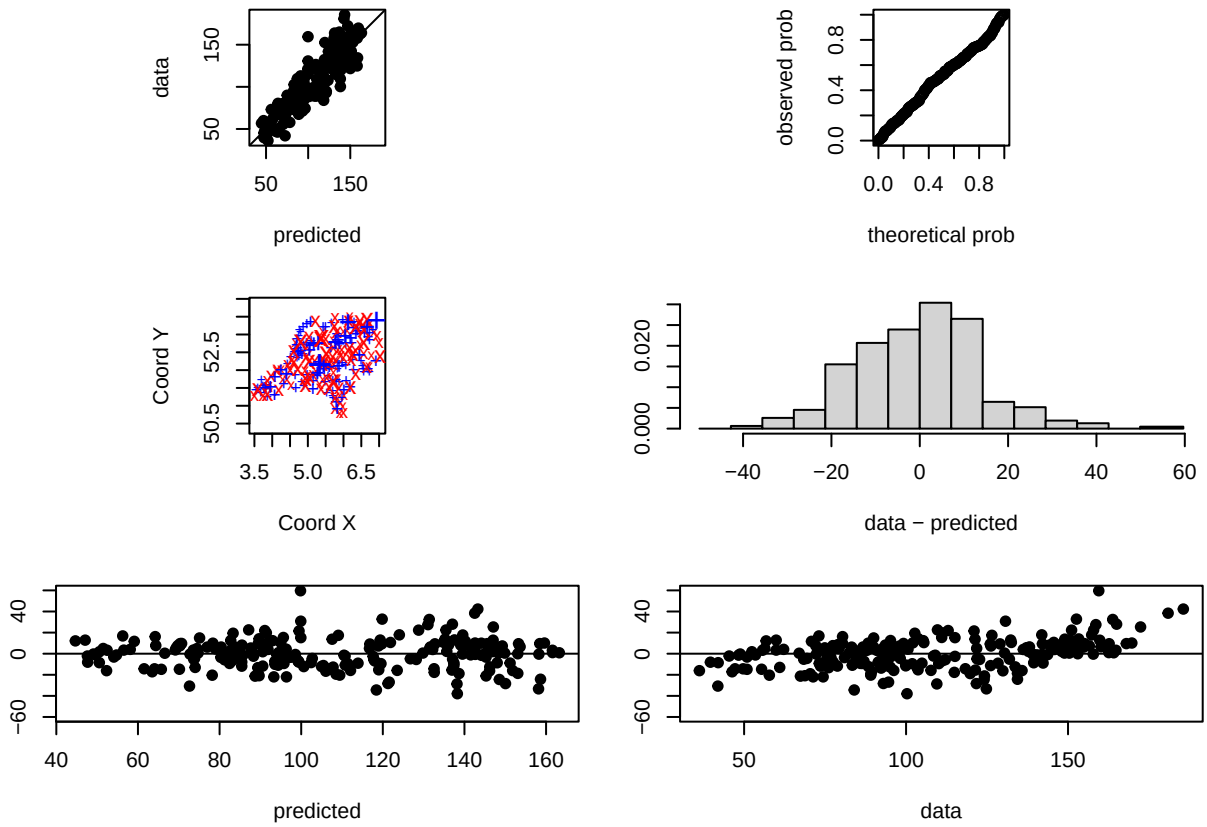

```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```




```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```




```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```




```
##                  Min.    1st Qu.    Median         Mean   3rd Qu.      Max.
## errors     -34.928225 -8.6999126 0.5227679 -0.030534243 8.4489395 56.996404
## std.errors  -2.486348 -0.6244408 0.0376028 -0.001103788 0.5955349  4.096564
##                    sd
## errors     13.9291599
## std.errors  0.9765343
```

```
##                  Min.    1st Qu.     Median          Mean   3rd Qu.      Max.
## errors     -37.616017 -9.3526088 0.83989772 -0.0198696993 8.6582696 59.303850
## std.errors  -2.545357 -0.6005366 0.05752065 -0.0006847676 0.5760121  4.071131
##                    sd
## errors     14.1788598
## std.errors  0.9555508
```

```
##                  Min.    1st Qu.     Median          Mean   3rd Qu.      Max.
## errors     -37.960171 -9.2509677 0.57617508 -0.0143638472 8.8631443 59.667751
## std.errors  -2.601369 -0.6239712 0.03992289 -0.0005097304 0.5964391  4.145316
##                    sd
## errors     14.2519891
## std.errors  0.9721234
```

From the validation plot, Power is a poor model. Its residuals do not follow normal distribution, and residuals are allocated by regions: higher residuals lie in higher latitude, while lower residuals lie in lower latitude. We can clearly see the red and blue regions in the error map.

The other three models are quite a good fit. The Leave-on-out residuals are quite Normal, and there is no strong patterns or systematic biases in the residuals. In all the three models, there is a relatively strong relationship between the fitted and true values (top left plot), although we do slightly underestimate the data values that are over 150. The residuals are reasonably Normal - closely following the QQ line (top right). There's no clear pattern in the spatial residuals, with blue and red locations (corresponding to the sign of the

residual) mostly randomly scattered (middle left). Histogram of errors show a reasonably normal distribution (middle right). In the bottom two plots, we can see that the errors of data (bottom right) above 150 tend to be positive - which confirms that the models underestimate data above 150 while errors of data under 50 tend to be negative - the models overestimate lower values. Errors by prediction scatter randomly.

Statistical summary of the errors and standard errors of the three models shows that, out of the three models, errors of model Matern 3/2 are closer together (ranging from -35 to 56 with sd = 13.85) while those of the other two are more distant.

For these reasons, Matern 3/2 is the best fit model out of these models.

Maximum Likelihood Model

Fit the maximum likelihood model

	model	trend	loglikehood	AIC
11	cubic	1st	-877.9548	1771.910
5	spherical	1st	-878.7807	1773.561
2	exponential	1st	-880.3243	1776.649
14	matern	1st	-880.3243	1776.649
12	cubic	2nd	-877.3556	1776.711
6	spherical	2nd	-878.1445	1778.289
8	circular	1st	-881.3759	1778.752
3	exponential	2nd	-879.8930	1781.786
15	matern	2nd	-879.8930	1781.786
9	circular	2nd	-881.0283	1784.057
4	spherical	cte	-891.0904	1794.181
1	exponential	cte	-891.9485	1795.897
13	matern	cte	-891.9485	1795.897
7	circular	cte	-892.9145	1797.829
10	cubic	cte	-893.0123	1798.025

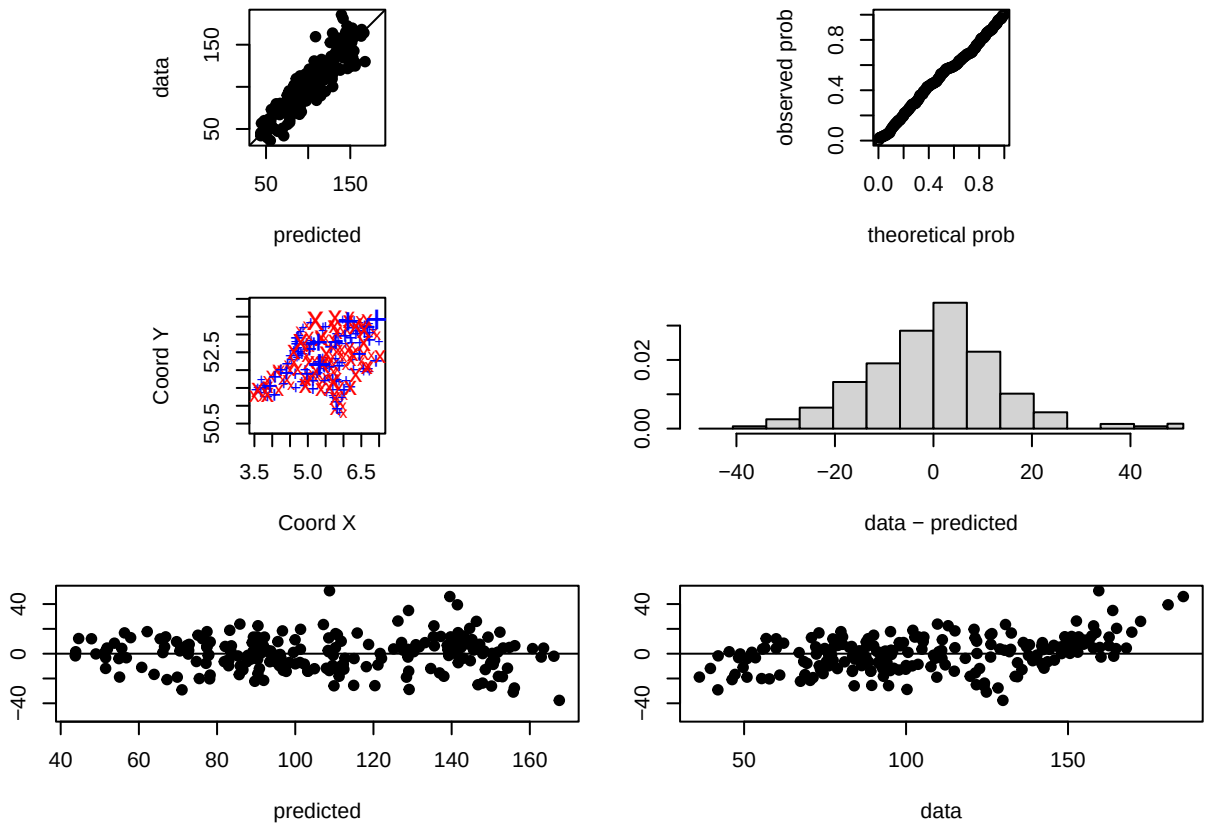
From the table results, it can be seen that models with the lowest scores of AIC and comparatively high maximised log-likelihood all have the mean be the first order polynomial on the coordinates. We will look into these four models: cubic, spherical, exponential and matern.

All these models estimate the estimate of lambda = 0.5, meaning that the transformation is the square root of the mean, while the nugget is estimated more than 1. Estimates of correlation length vary from 0.1 to 0.5.

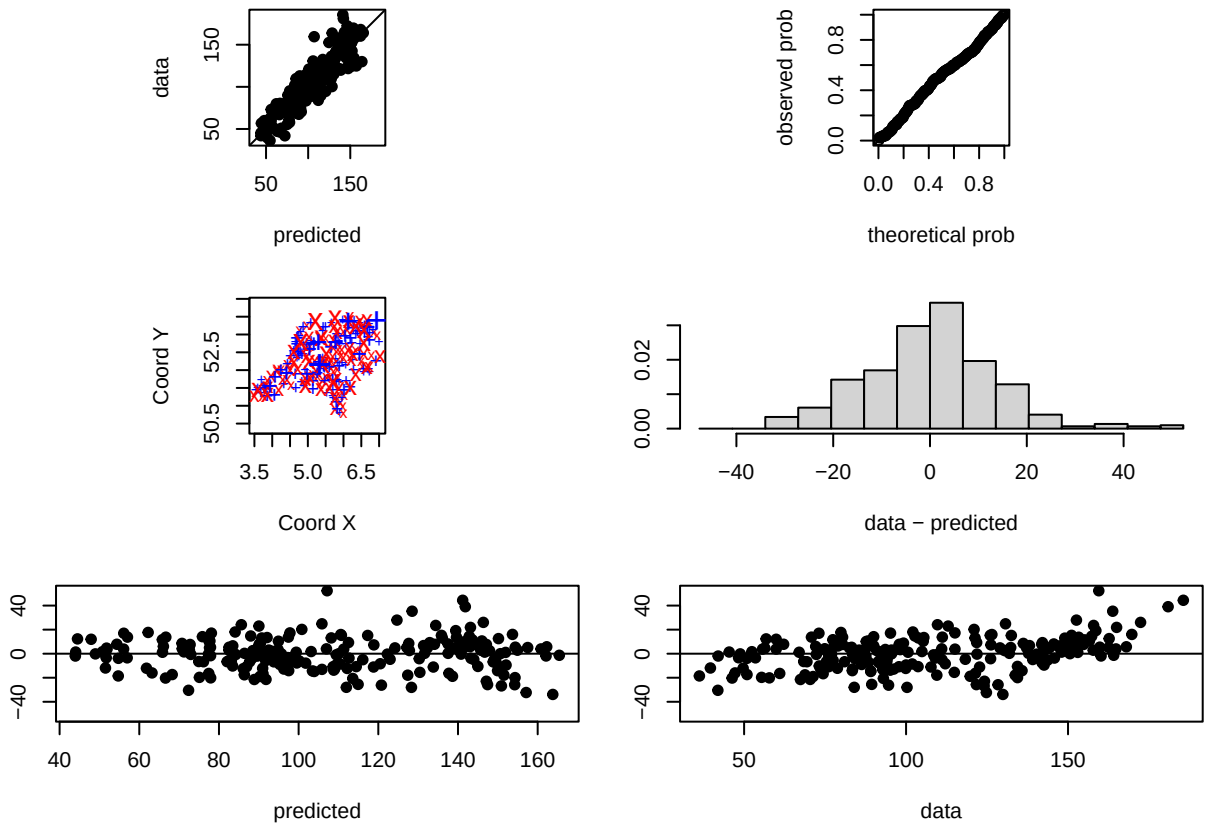
Betas indicate the coefficients of the mean function, with beta1 and beta2 respectively coefficients of the spatial coordinates. All models result in high beta2 (~ from 4 to 5), confirming the positively linear relationship between data values and latitude mentioned in part a. Beta1 are estimated to be around -1, showing a slightly negative relationship with longitude.

We validate these 5 models (validation results are in full version)

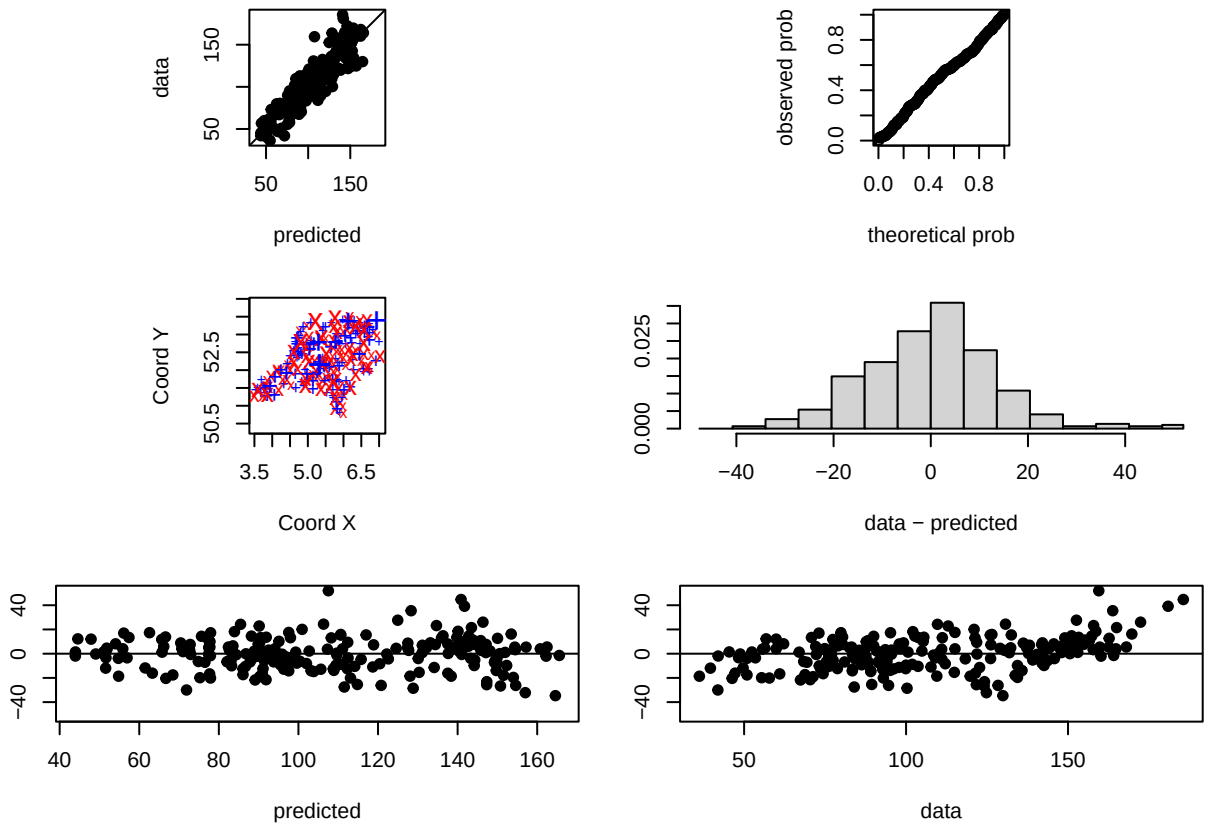
```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```



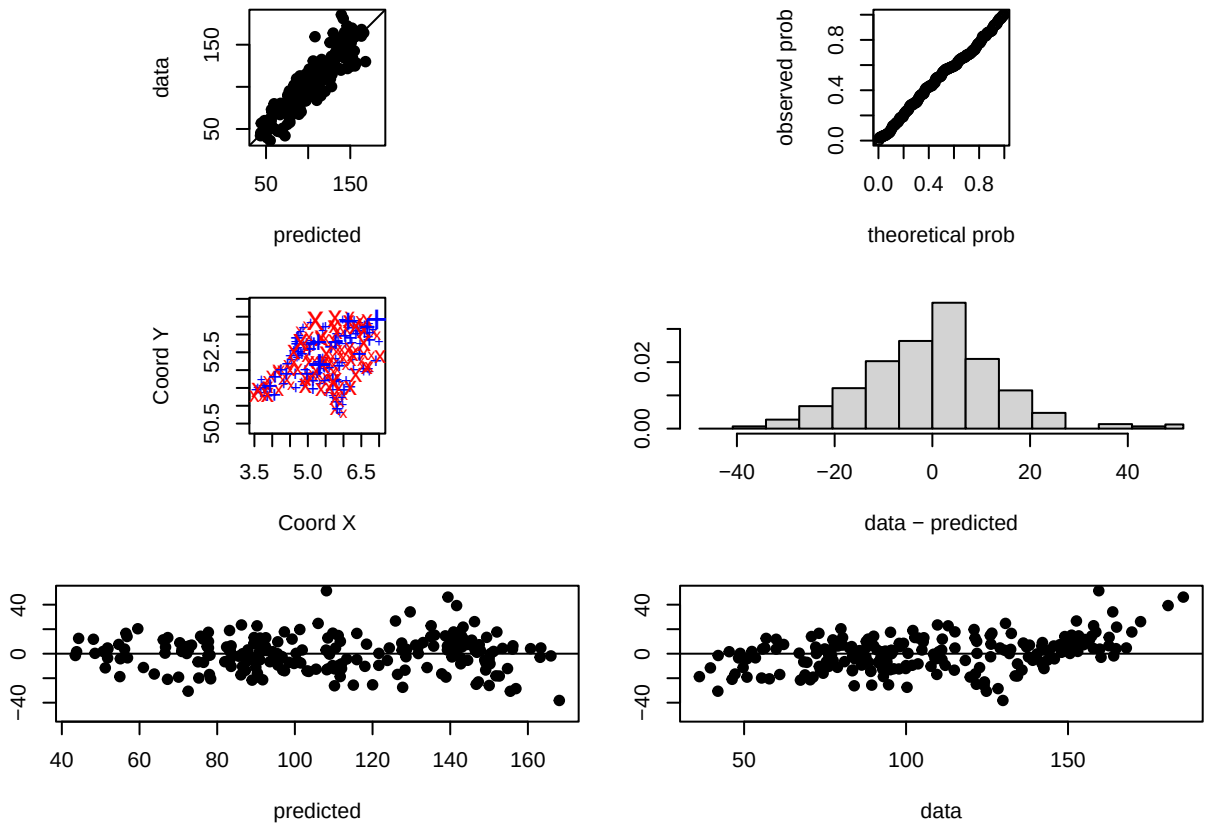

```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```



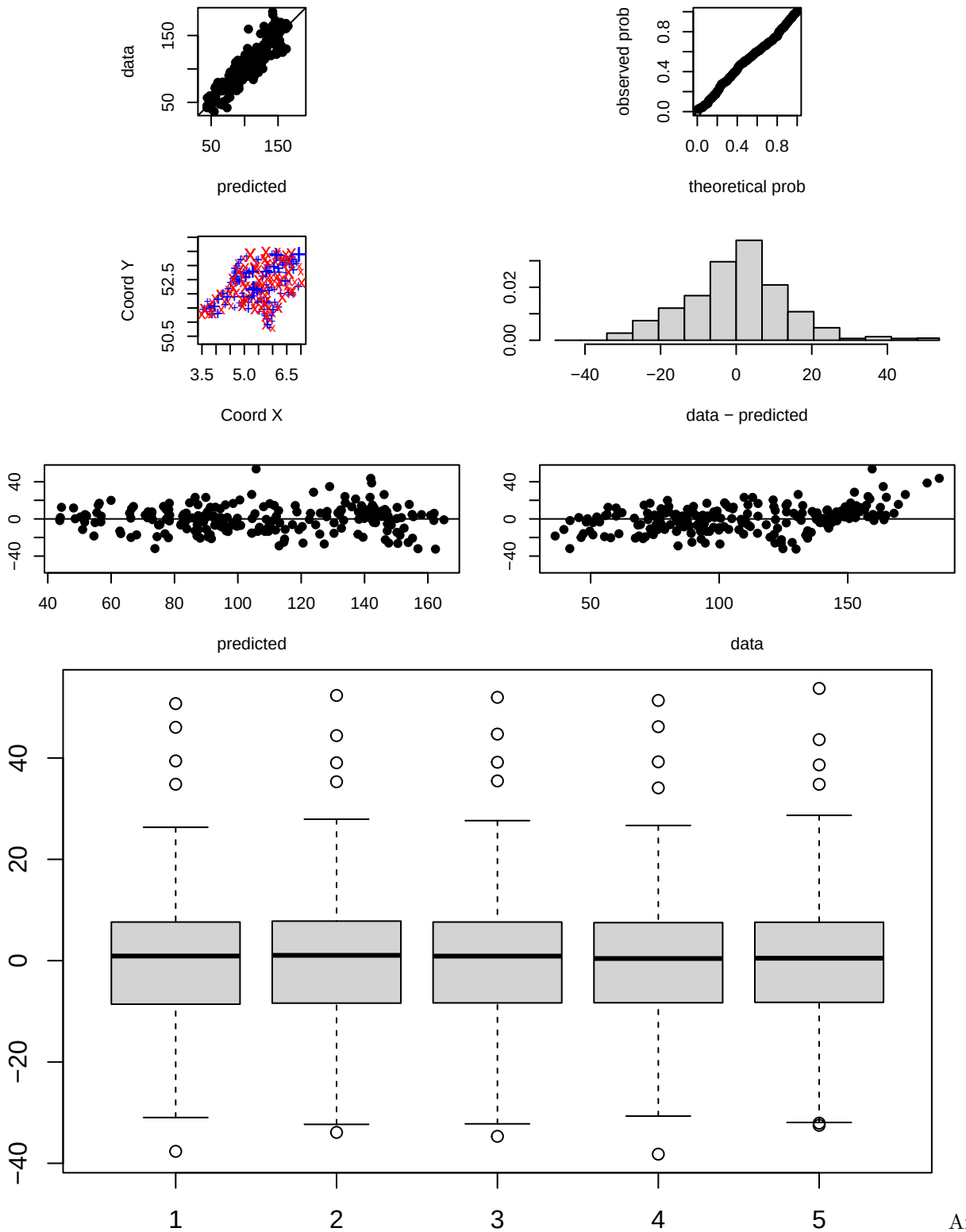

```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```




```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```




```
## xvalid: number of data locations       = 217
## xvalid: number of validation locations = 217
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```

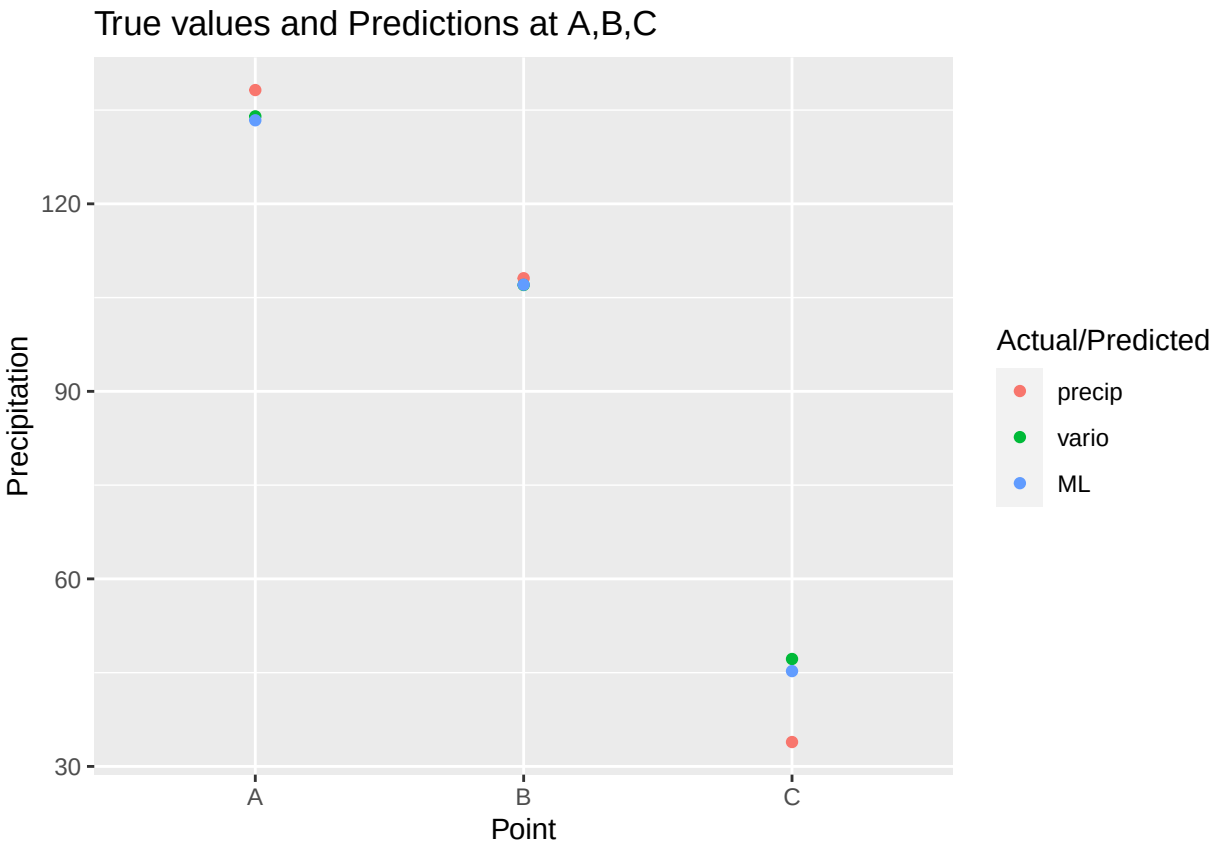


After validation, all models look a good fit to the data. All 6 plots of 5 models are comparatively identical. The predicted values scatter on both sides of the "data line" (top left) and the probability follow strictly the QQ line, showing a good Normal distribution (top right). Middle left map show a mixture of blue and red points over the region, and middle right shows histogram of error which looks normal. The bottom plots illustrate errors against data and predicted values, with no strong patterns to be found. However, as in the Variogram

model, all models seems to overestimate data under 50 and underestimate data above 150. This result is acceptable, and as we have plugged in most optional arguments, we will try improving models by other methods in later parts.

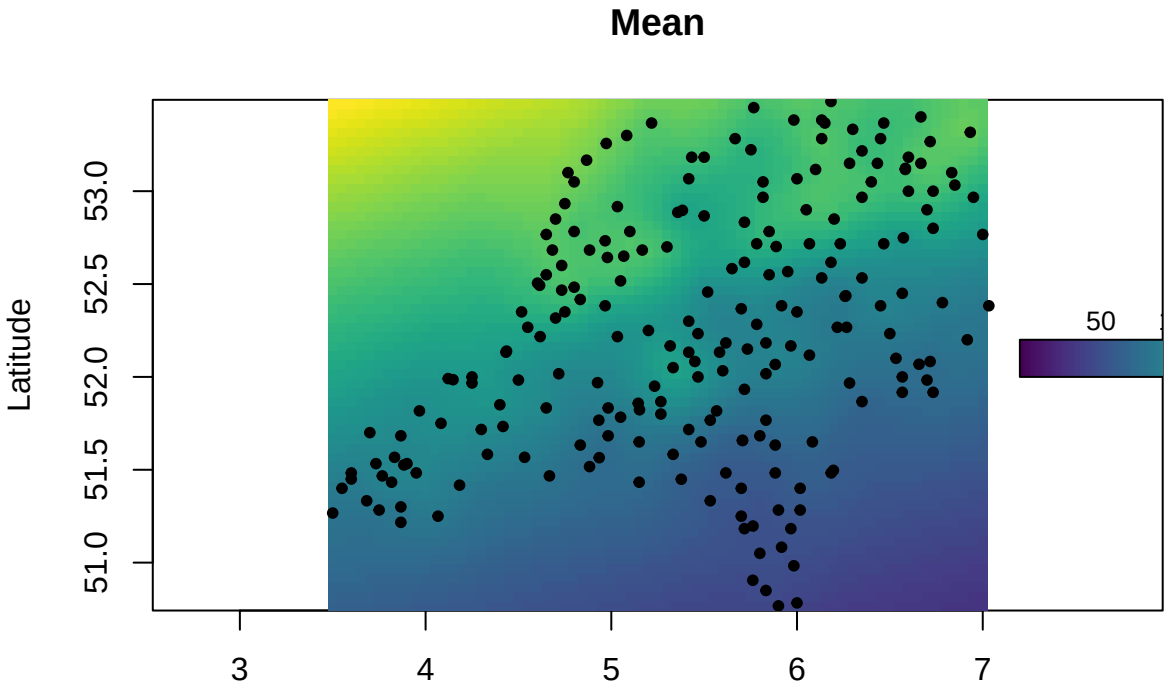
The fourth model, which is spherical, has the shortest interquartile range of errors, thus, could be considered the best models in maximum likelihood.

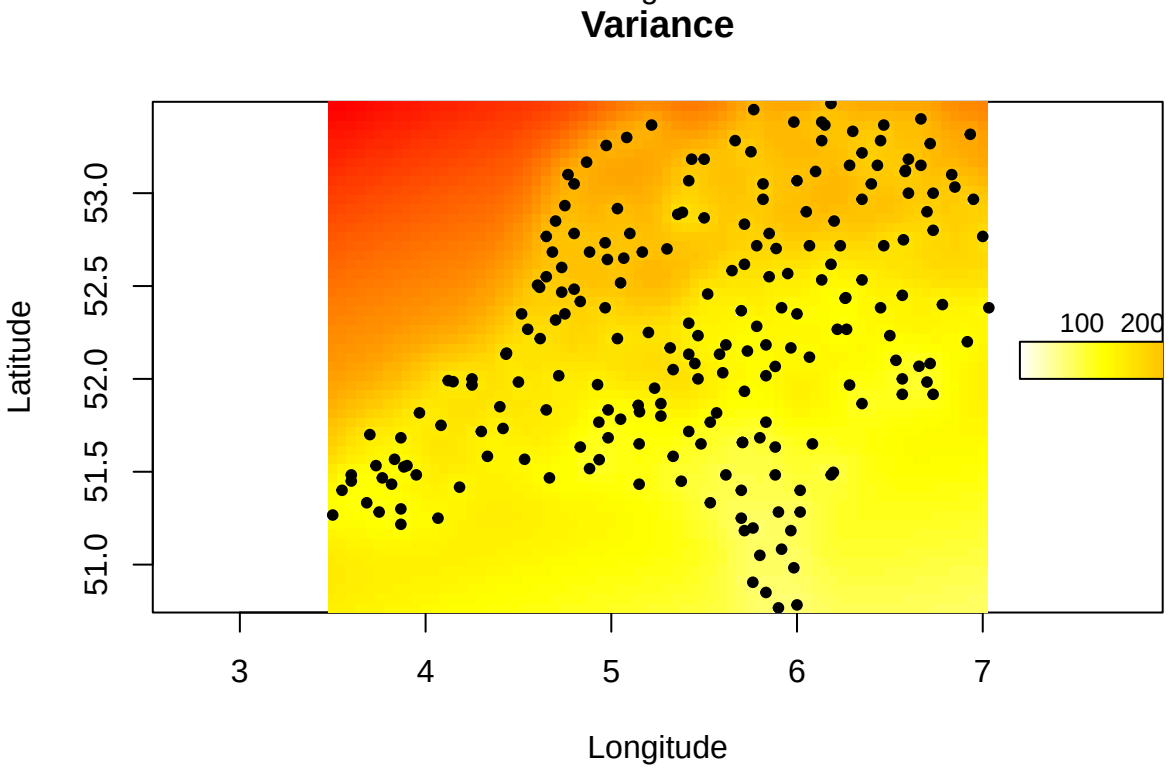
Predict precipitation at A, B, C



Compare predictions and true values The actual values (red points) and predicted values (green by variogram and blue by ML) are shown in the above plot. The actual values and predictions are quite close at A and B, however, both models overestimated precipitation at C. Both predicted C to have around 45, however, observed figure at C was just above 30. At all three locations, both estimation methods produce closely similar prediction (in terms of the mean prediction).

Plot the mean and variance from maximum likelihood





Fit a Bayesian model using discrete priors.

From the estimates of multiple maximum likelihood models, we recall that the lambda is around 0.5, the nugget is more than 1 and the phi ranges from 0.1 to 0.5. Besides, from the previous analysis, we can have a

reasonable assumption that the mean is the first order polynomial of coordinations and the covariance matrix is spherical.

We can use estimates of parameters from maximum likelihood spherical as reference.

```
##            status    values
## beta0   estimated -233.5359
## beta1   estimated   -1.4849
## beta2   estimated    4.9883
## tausq   estimated    1.3076
## sigmasq estimated    1.1792
## phi     estimated    0.5180
## kappa       fixed    0.5000
## psiA        fixed    0.0000
## psiR        fixed    1.0000
## lambda  estimated    0.5114
```

Build Bayes GP with and without nugget

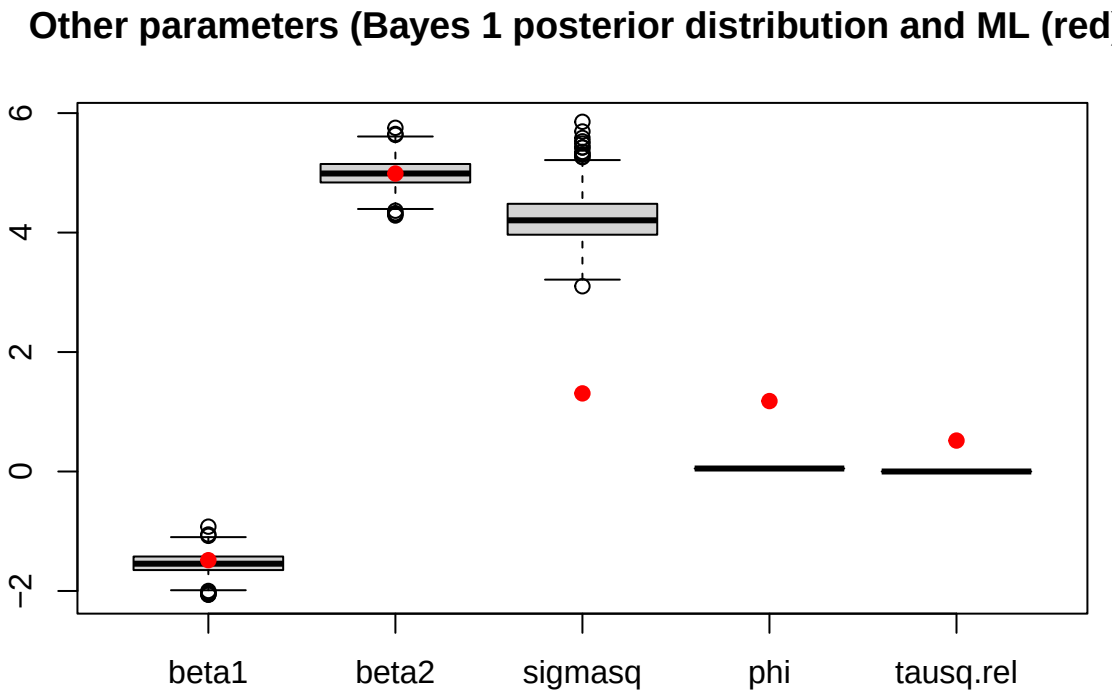
We will summarise our posterior distributions, compare them against each other and compare them with earlier estimates.

From the four plots below, the posterior of the model 2 (with nugget) is quite close with the parameter estimates from the maximum likelihood model, especially betas parameters and variance. For phi and nugget, maximum likelihood estimates are out of interquartile range of distributions from Bayes (with nugget), however, they are not too far away.

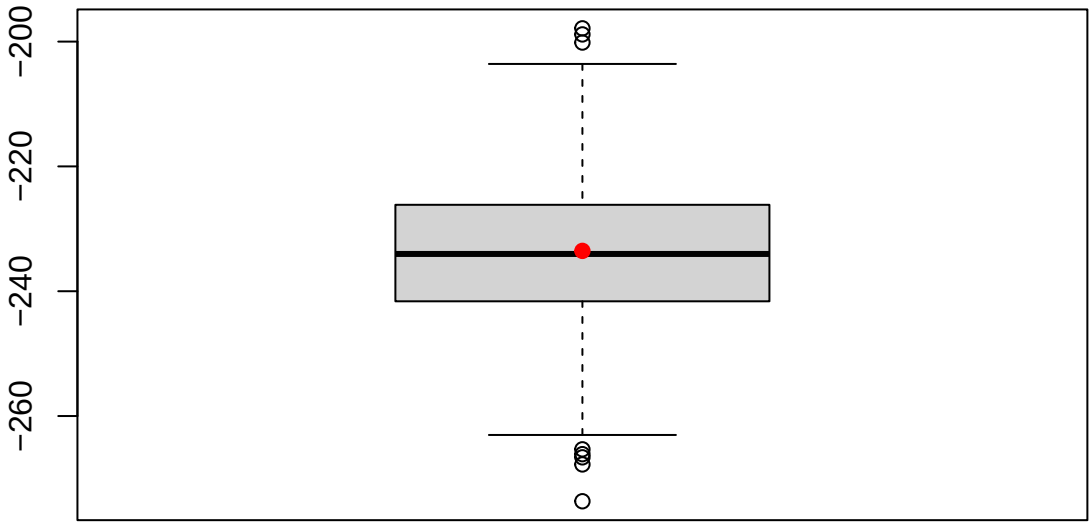
Meanwhile, the posterior distribution of the model 1 (without nugget) only closely matches with the maximum likelihood estimates in beta parameters. Its variance and correlation length are far from the estimates by ML.

Comparing Bayesian with and without nugget, we can see that the Bayesian 1 without nugget produce much higher estimates of variance (mean around 4) than the Bayesian 2 with nugget (mean around 1). Distribution of correlation length and nugget by Bayesian 2 are sensible, quite close with our earlier estimates. Both Bayesian models estimate quite similar betas parameters.

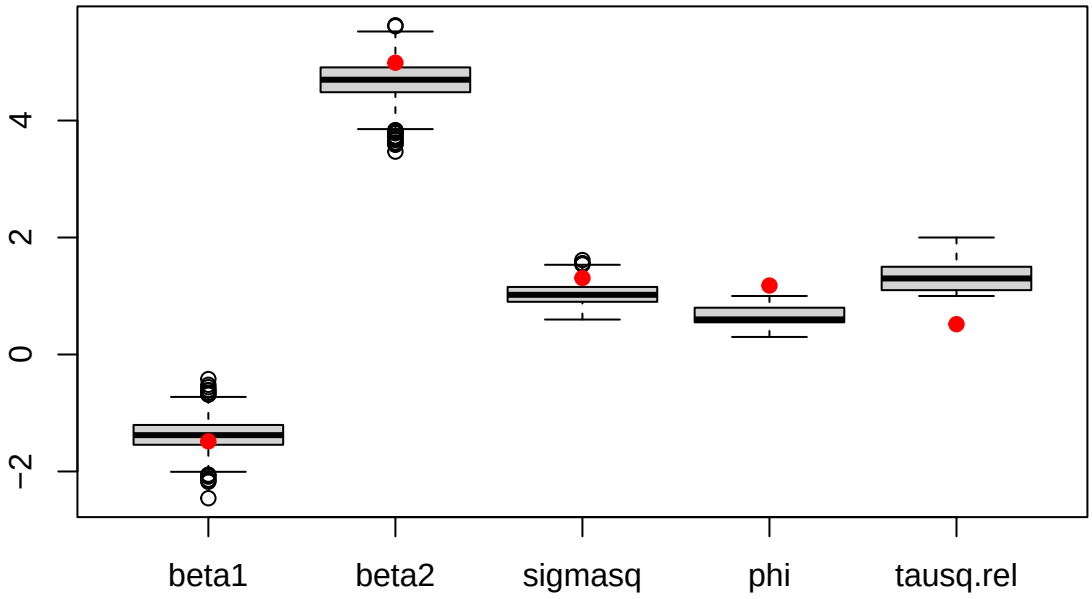
We will use cross-validation to validate both models

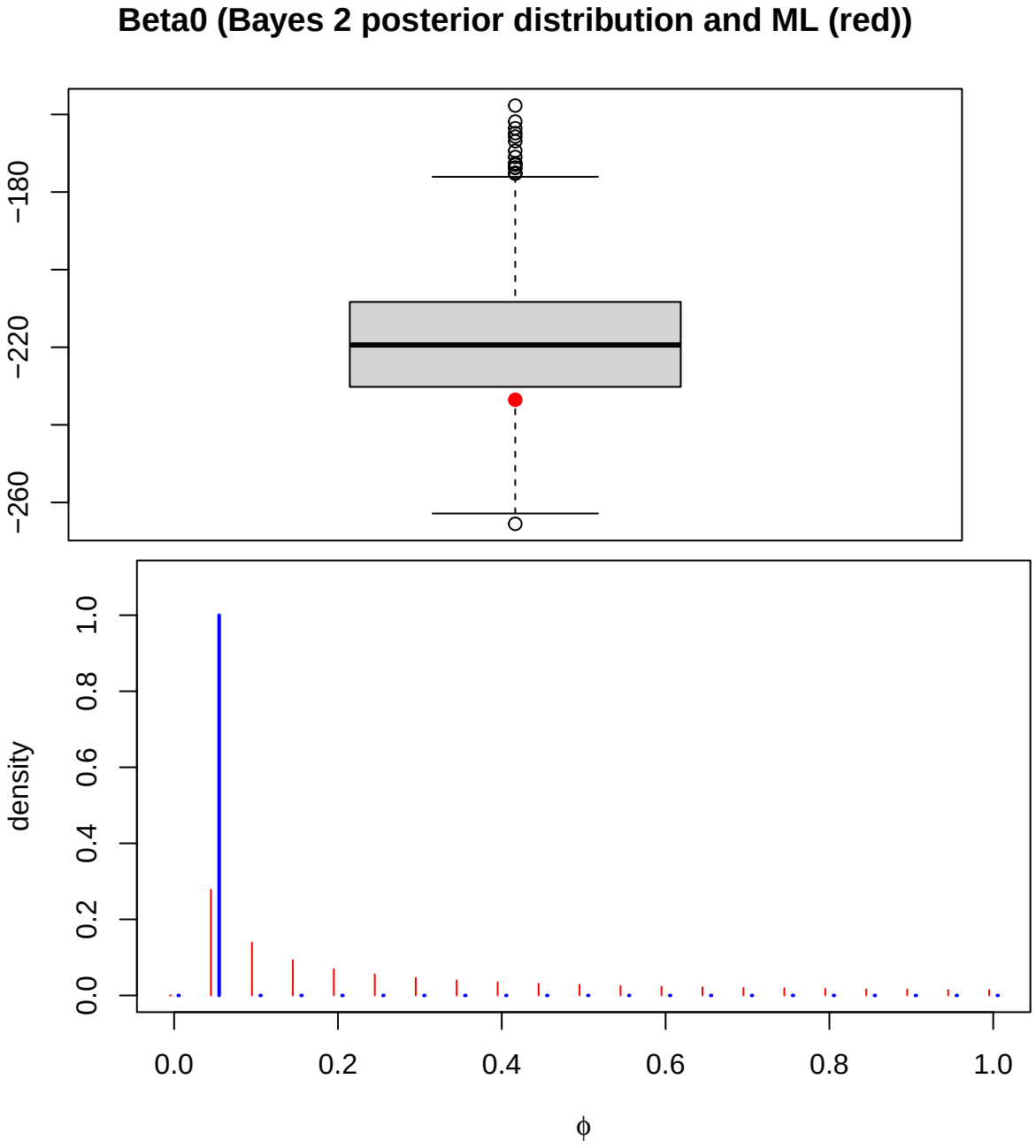


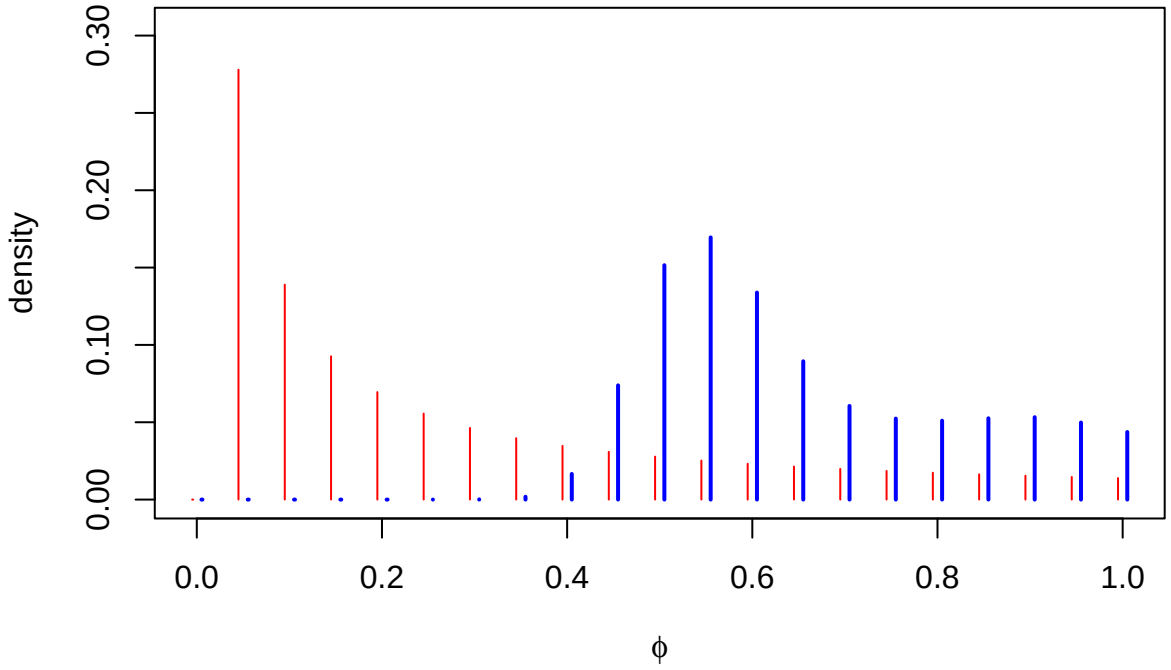
Beta0 (Bayes 1 posterior distribution and ML (red))

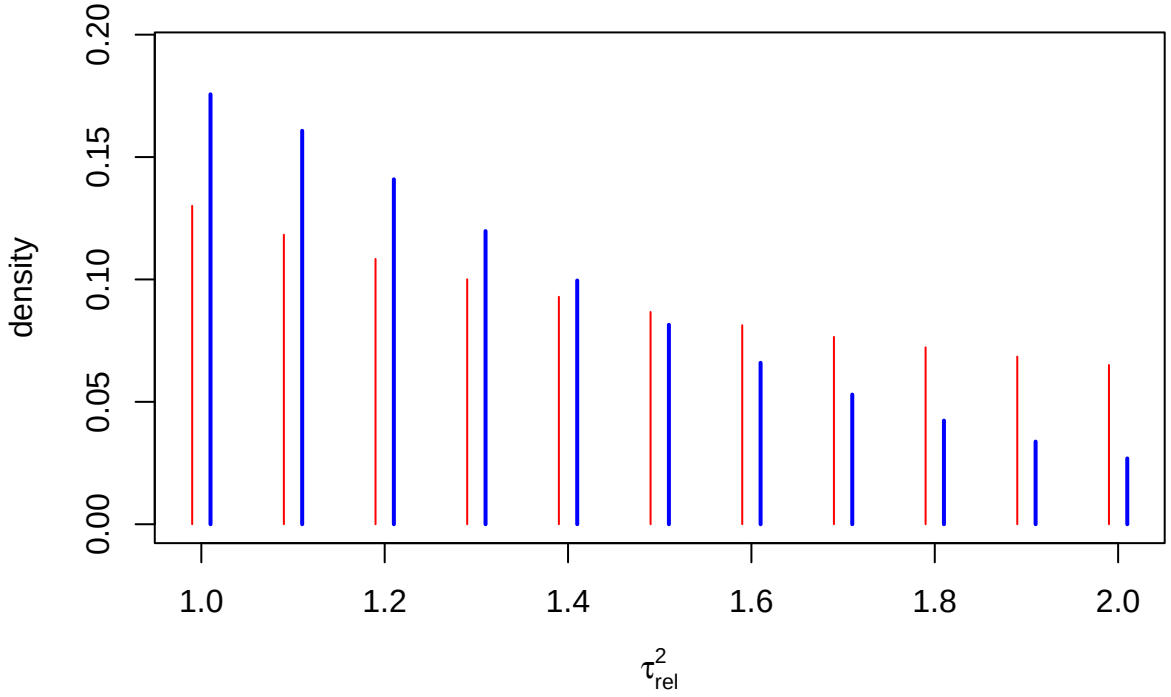


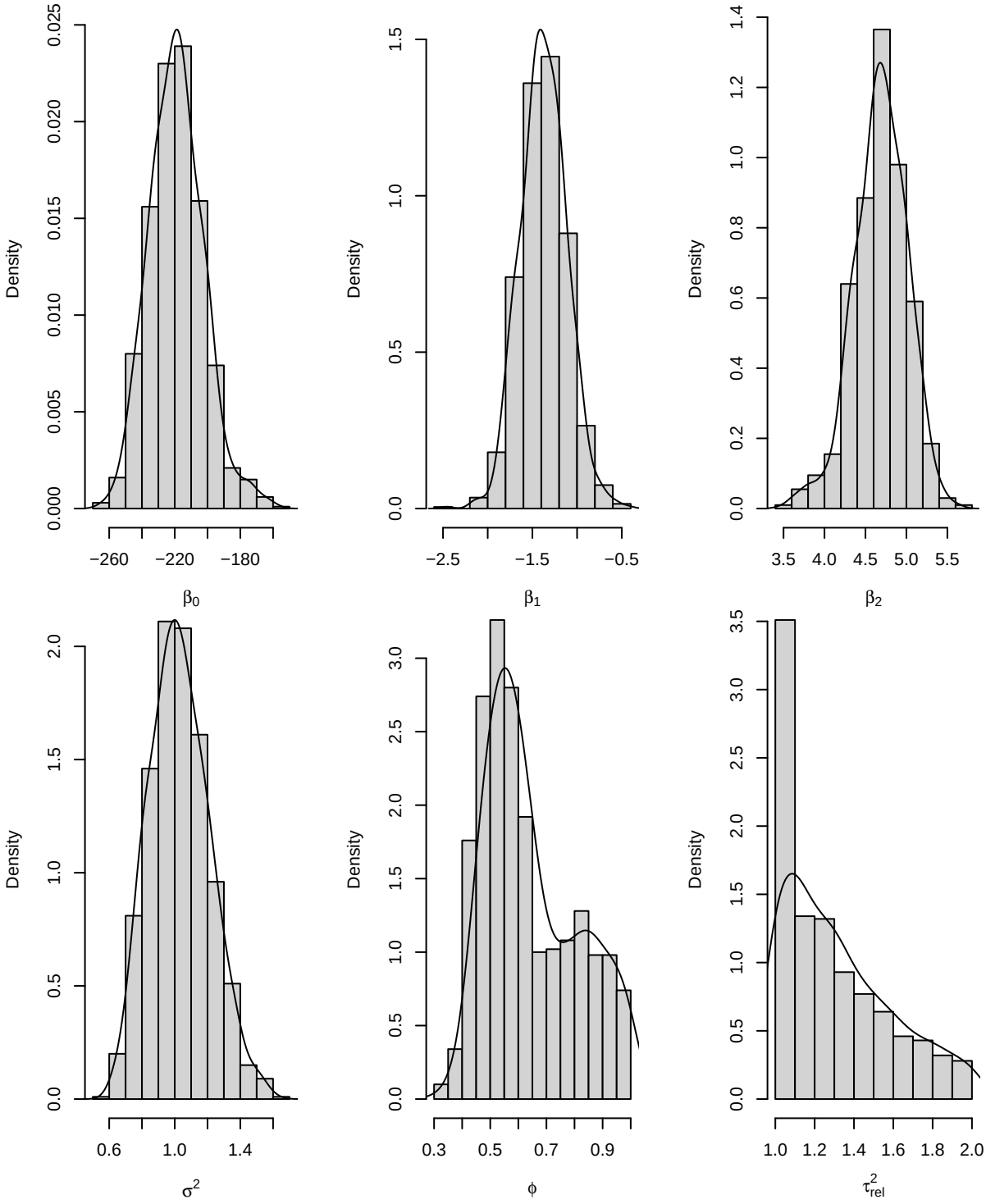
Other parameters (Bayes 2 posterior distribution and ML (red))



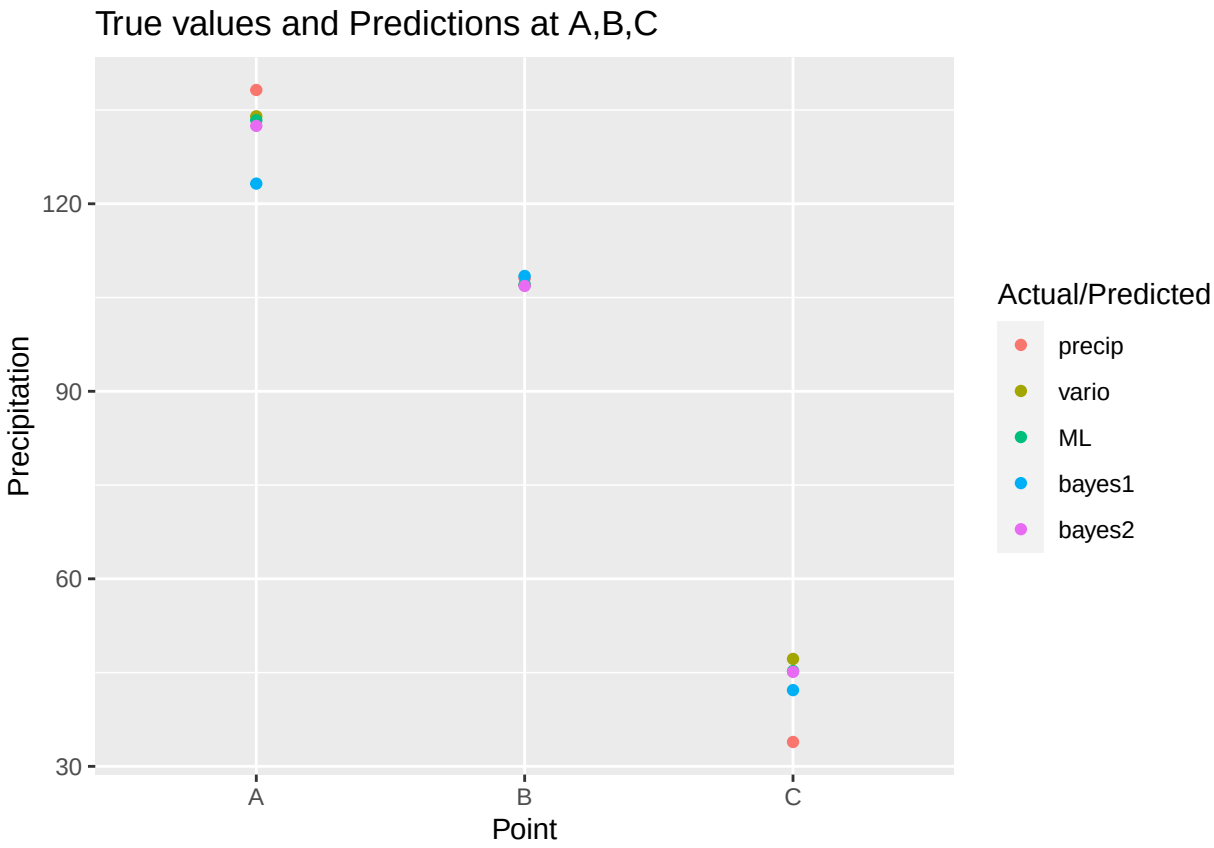

Beta0 (Bayes 2 posterior distribution and ML (red))
−180
−220
−260
density
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ϕ


density
0.00
0.10
0.20
0.30
0.0
0.2
0.4
0.6
0.8
1.0
ϕ


density
0.00
0.05
0.10
0.15
0.20
1.0
1.2
1.4
1.6
1.8
2.0
τ^2_{rel}



For each model, produce predictions for locations A, B and C.



As can be seen from the plot, Bayesian without nugget poorly predicts precipitation at A, but performs similarly with other models at B and C, especially at C, its estimate is closest to the actual value. Meanwhile, Bayesian 2 performs similarly with the other two models at predicting values at A, B and C.