Estimating Daily PM2.5 Concentration by Spatial Statistical Downscaling with INLA

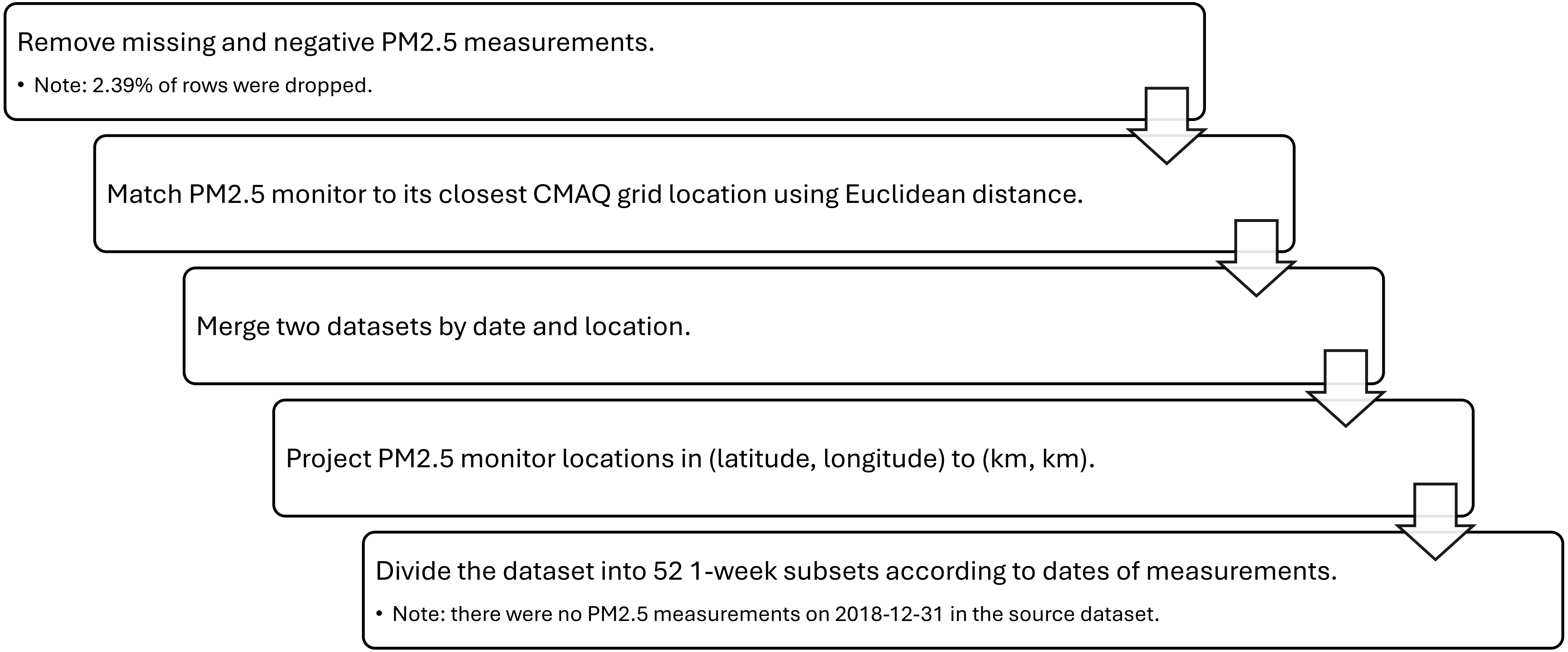
Mar 13, 2024

Data

- Daily PM2.5 concentrations in the contiguous US in 2018
  - Variables: monitor site id, latitude and longitude of monitor stations, PM2.5 in $\mu g/m^3$, date, coordinate reference system

- Simulated CMAQ PM2.5 concentrations by National Center for Atmospheric Research on a 12km*12km grid
  - Variables: simulated PM2.5, grid centroid latitude and longitude, date

Processing Steps



Remove missing and negative PM2.5 measurements.

- Note: 2.39% of rows were dropped.

Match PM2.5 monitor to its closest CMAQ grid location using Euclidean distance.

Merge two datasets by date and location.

Project PM2.5 monitor locations in (latitude, longitude) to (km, km).

Divide the dataset into 52 1-week subsets according to dates of measurements.

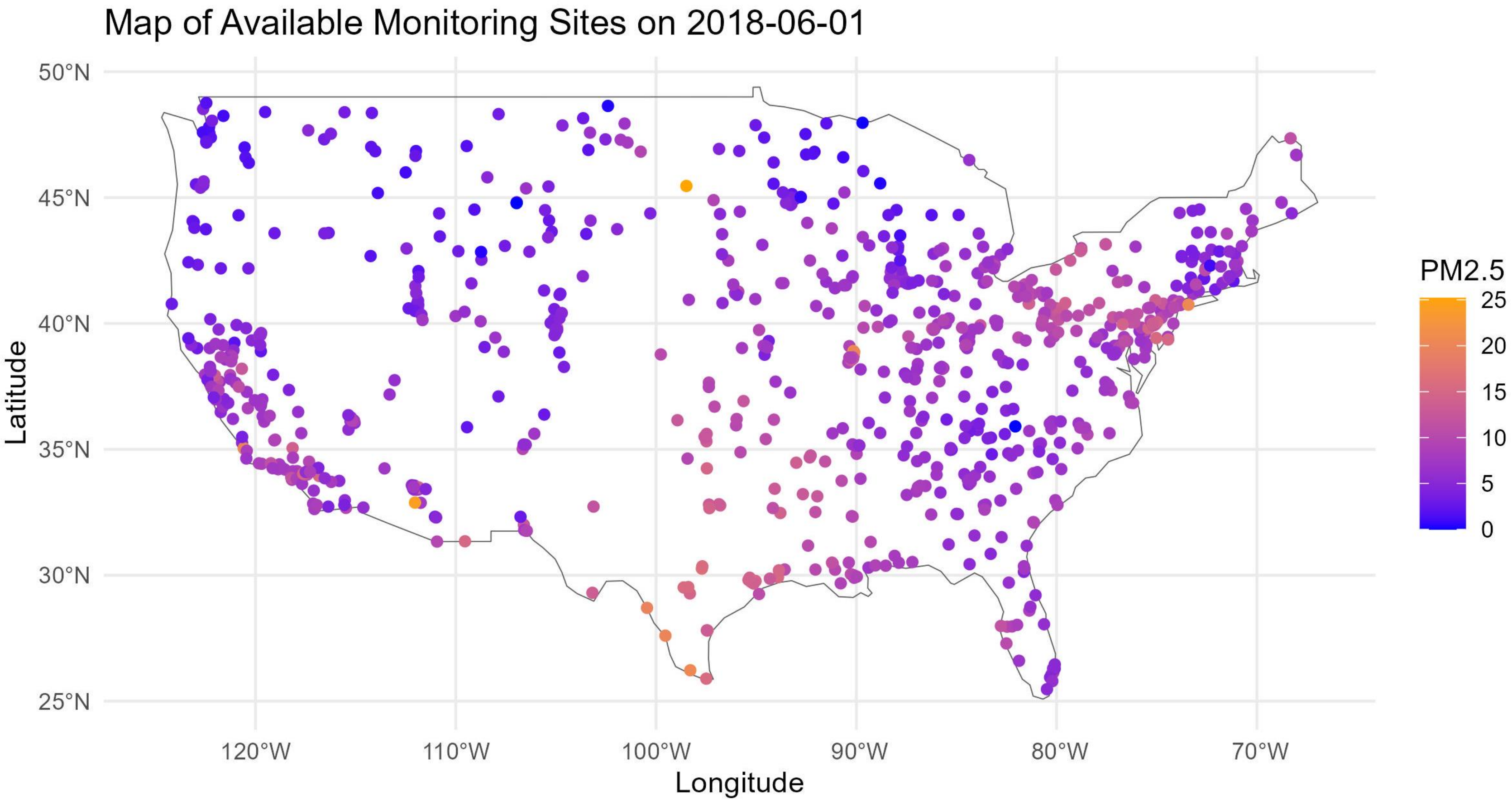
- Note: there were no PM2.5 measurements on 2018-12-31 in the source dataset.

Descriptive Table

Characteristic	**Overall**, N = 235,402	**Q1**, N = 57,494	**Q2**, N = 58,285	**Q3**, N = 59,741	**Q4**, N = 59,882
Daily PM2.5					
Mean (SD)	8.3 (7.3)	8.0 (6.2)	7.5 (4.1)	9.6 (8.1)	8.3 (9.3)
Range	0.0, 411.7	0.0, 318.8	0.0, 167.8	0.0, 261.0	0.0, 411.7
Simulated CMAQ PM2.5					
Mean (SD)	8.6 (5.6)	8.4 (5.8)	7.5 (4.1)	9.7 (5.7)	8.7 (6.4)
Range	0.0, 75.5	0.1, 54.3	0.1, 44.4	0.1, 75.5	0.0, 74.4
Monitors					
Count	969	936	937	930	939

Q1: January – March
Q2: April – June
Q3: July – September
Q4: October – December

An Example Plot



R-INLA

- The integrated nested Laplace approximation (INLA) is a method for **approximate Bayesian inference**.

- It has established itself as an alternative to other methods such as Markov chain Monte Carlo because of its **speed** and **ease of use** via the R-INLA package.

- Install instruction

Modeling PM2.5: SPDE

- Stochastic Partial Differential Equation (SPDE) approach approximates the continuous Gaussian field as a discrete Gaussian Markov random field. Predictions utilize a matrix that maps **a finite set** of observed locations to a set of unobserved locations.

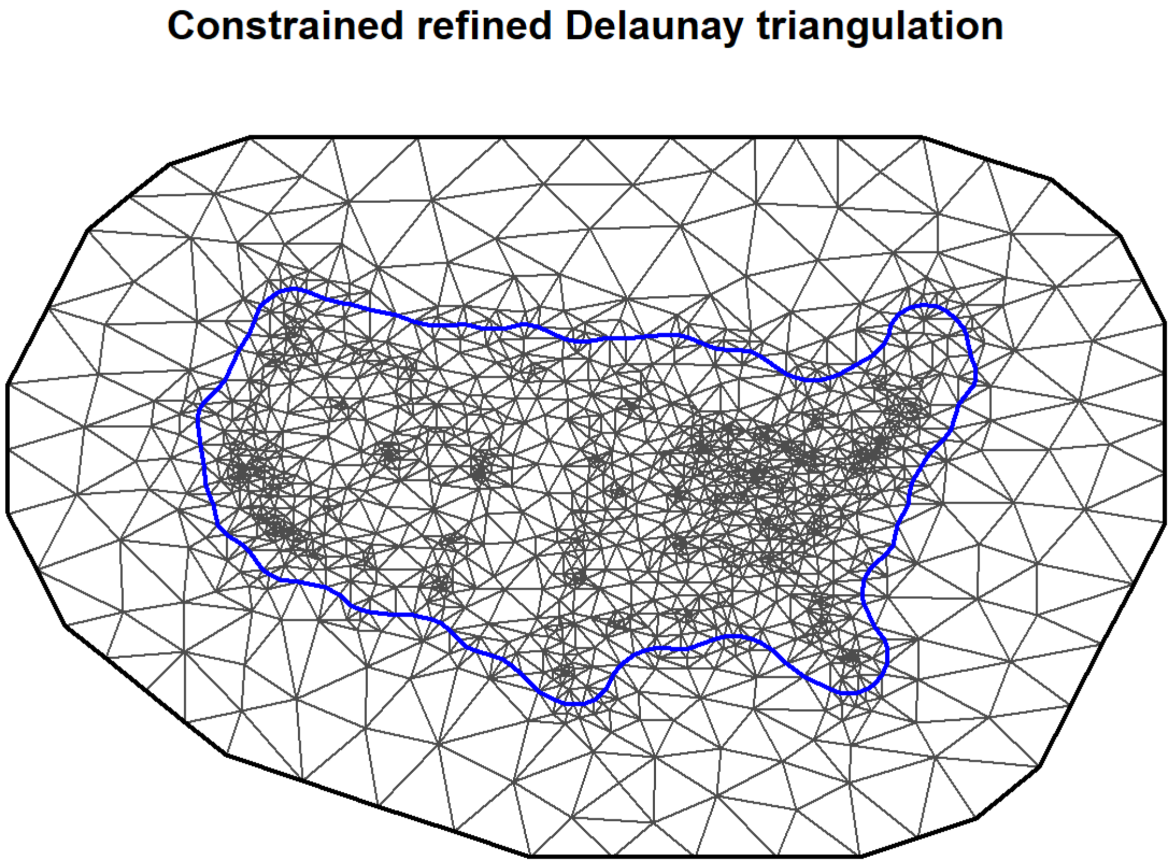
- SPDE essentials
    - Mesh
    - Range parameter ρ
    - Marginal standard deviation σ

- Prior function:

$$\pi(\rho,\sigma)=\frac{\mathrm{d}\lambda_\rho}{2}\rho^{-1-d/2}\exp\left(-\lambda_\rho\rho^{-\frac{d}{2}}\right)\lambda_\sigma\exp(-\lambda_\sigma\sigma)$$

where λ_ρ and λ_σ are hyperparameters

Mesh Example



Constrained refined Delaunay triangulation

Larger triangles are specified close to the boundary to alleviate boundary effect.

Models

Given n PM2.5 observations $y_i, i = 1, \dots, n$, at locations s_i and date $t_i = 1, \dots, 364$, the models can be defined as:

$$y_i | \mu_i, \sigma_\epsilon \sim N(\mu_i, \sigma_\epsilon^2)$$
$$\mu_i = .$$
$$u | \sigma, \rho \sim GP(0, \Sigma)$$

where σ_ϵ is the iid Gaussian noise and μ_i is defined differently in four sets of models:

Model	Formula
1.1	$\mu_i = \beta_0(s_i) + u(s_i)$
1.2	$\mu_i = \beta_0(s_i, t_i) + w(s_i, t_i)$
2.1	$\mu_i = \beta_0(s_i) + \beta_1 X_i + u(s_i)$
2.2	$\mu_i = \beta_0(s_i, t_i) + \beta_1 X_i + w(s_i, t_i)$
3.1	$\mu_i = \beta_0(s_i) + \beta_1(s_i) X_i + u(s_i)$
3.2	$\mu_i = \beta_0(s_i, t_i) + \beta_1(s_i) X_i + w(s_i, t_i)$
4.1	$\mu_i = \beta_0(s_i) + \beta_1(s_i, t_i) X_i + u(s_i)$
4.2	$\mu_i = \beta_0(s_i, t_i) + \beta_1(s_i, t_i) X_i + w(s_i, t_i)$

- y_i (outcome) denotes the PM2.5 observations.
- X_i (exposure) denotes the simulated CMAQ PM2.5.

- $\beta_0(s_i)$ is the intercept as a spatially varying coefficient.
- $\beta_0(s_i, t_i)$ is the intercept as a space-time varying coefficient.

- β_1 is the coefficient of the exposure.
- $\beta_1(s_i)$ is the spatially varying coefficient of the exposure.
- $\beta_1(s_i, t_i)$ is the space-time varying coefficient of the exposure.

- $u(s_i)$ is a spatial random effect that follows a zero-mean Gaussian process with Matérn covariance function.
- $w(s_i, t_i)$ is a random effect that changes in time with first order autoregressive dynamics and spatially correlated innovations.

$$w(s_i, t_i) = aw(s_i, t_i - 1) + u(s_i, t_i)$$

Models Continue

- **Spatial random effect**

$$u(s_i) \sim GP(0, \Sigma)$$

with Matérn covariance function.

- **Space-time random effect**

$$w(s_i, t_i) = aw(s_i, t_i - 1) + u(s_i, t_i)$$
$$u(s_i, t_i) \sim GP(0, \Sigma)$$

For $|a| < 1$ and $w(s_i, 1)$ follows a stationary distribution of a first-order autoregressive process, namely, $N\left(0, \sigma_u^2/(1 - a^2)\right)$. Each $u(s_i, t_i)$ follows a zero-mean Gaussian distribution temporally independent but spatially dependent at each time with Matérn covariance function.

Source: Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman and Hall/CRC.

Matérn Covariance

- Matérn covariance function for a spatial random effect $u(s_i)$:

$$cov\left(u(s_i), u(s_j)\right) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\kappa||s_i - s_j||\right)^{\nu} K_\nu(\kappa||s_i - s_j||)$$

where $K_\nu(\cdot)$ is the modified Bessel function of second kind and order $\nu > 0$. ν is the smoothness parameter, σ^2 denotes the variance, and $\kappa > 0$ is related to the range as $\rho = \sqrt{8\nu/\kappa}$.

- Matérn covariance function for a space-time random effect $u(s_i, t_i)$:
  - When $t_i = t_j$,

$$cov\left(u(s_i, t_i), u(s_j, t_j)\right) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\kappa||s_i - s_j||\right)^{\nu} K_\nu(\kappa||s_i - s_j||)$$

  - When $t_i \neq t_j$,

$$cov\left(u(s_i, t_i), u(s_j, t_j)\right) = 0$$

since in our previous convention that the time component is independent.

Source: Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman and Hall/CRC.

Priors

Matérn SPDE is used with penalized complexity (PC) priors.

- **Range**

$$p(\rho < 20) = 0.05$$

  - The probability that the **range** is smaller than 20km is very small – around 0.05. And the **range** of the process is the distance at which the spatial correlation is close to 0.1.

- **Marginal standard deviation**

$$p(\sigma > 22) = 0.01$$

  - The probability that the variability (as **marginal standard deviation**) of PM2.5 data is greater than 22 $\mu g/m^3$ is very small – around 0.01.

Source: Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman and Hall/CRC.

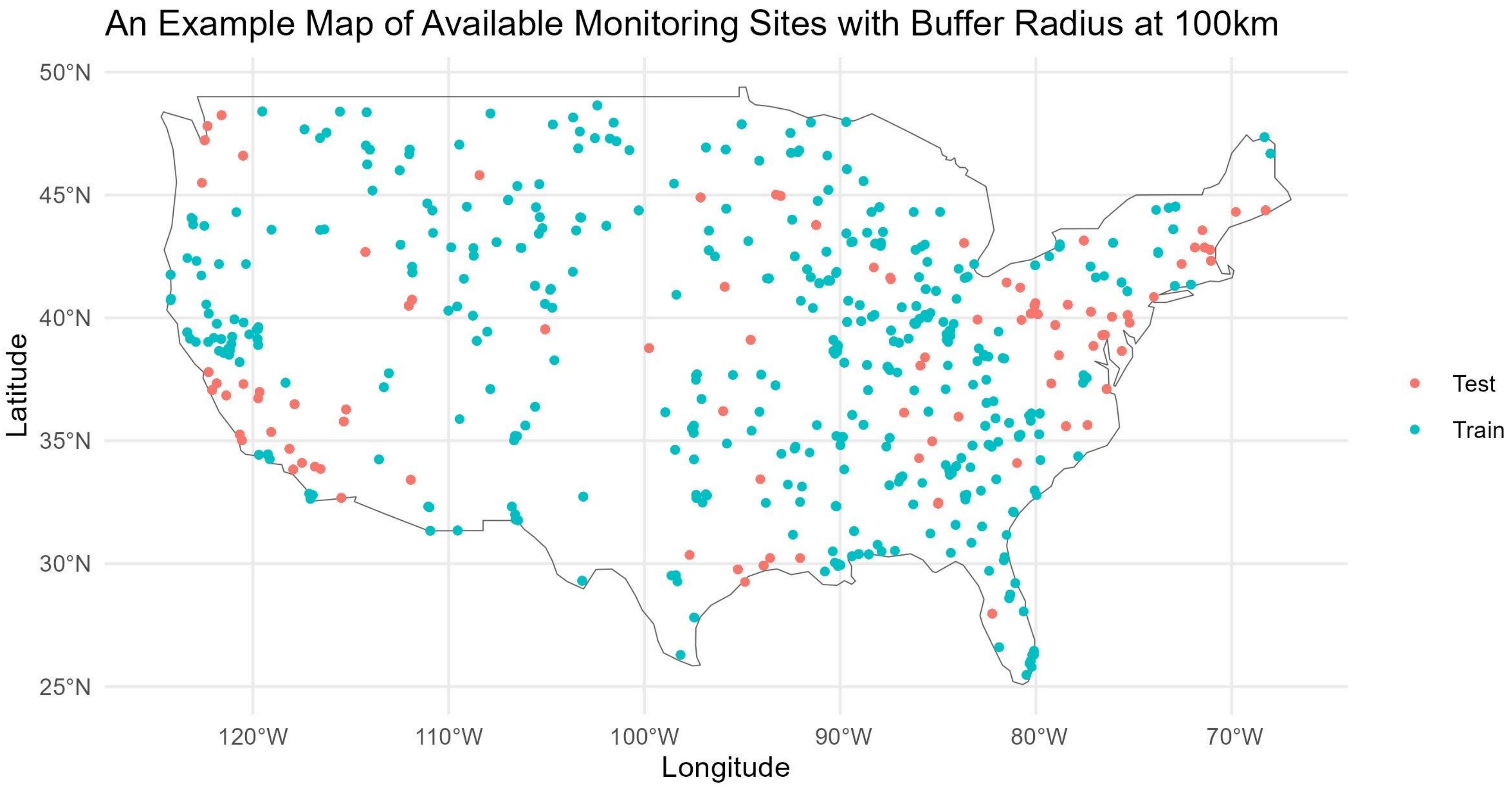
Model Selection

- 10-fold Spatial Cross-validation is performed to select an optimal model among the 8 models that we previously defined, based on the **root mean square error** (smaller is better) and the **coverage** (larger is better) of 95% prediction intervals.

- To obtain folds, only locations are sampled, meaning that the test data are every day's PM2.5 measurements at sampled sites in the 1-week domain.

- Normally train and test data in CV are independent, but we cannot assume that for spatial data.
    - Solution: exclude disks around test data to reduce dependence.

Plot of Train and Test Data



RMSE by Quarters

- Model 2.1 and 2.2 have the smallest RMSE overall and in all quarters.
- PM2.5 are predicted the most accurately in Q2 (Apr - Jun) by all models.

Root Mean Square Error					
	Q1	**Q2**	**Q3**	**Q4**	**Overall**
Model 1.1	5.96	4.08	7.68	9.04	6.97
Model 1.2	5.96	4.08	7.68	9.04	6.97
Model 2.1	5.31	3.59	6.67	7.92	6.11
Model 2.2	5.31	3.59	6.67	7.92	6.11
Model 3.1	5.79	4.01	7.24	8.05	6.48
Model 3.2	5.65	3.94	6.95	7.53	6.19
Model 4.1	5.85	4.04	7.43	8.49	6.69
Model 4.2	5.75	4	7.22	8.03	6.45

No buffer / radius 0km

Root Mean Square Error					
	Q1	**Q2**	**Q3**	**Q4**	**Overall**
Model 1.1	6.02	4.11	7.73	9.15	7.04
Model 1.2	6.02	4.11	7.73	9.15	7.04
Model 2.1	5.38	3.62	6.82	8.19	6.26
Model 2.2	5.38	3.62	6.82	8.19	6.26
Model 3.1	5.98	4.09	7.52	8.96	6.9
Model 3.2	5.94	4.06	7.35	8.79	6.79
Model 4.1	6	4.1	7.63	9.06	6.97
Model 4.2	5.97	4.08	7.53	8.97	6.9

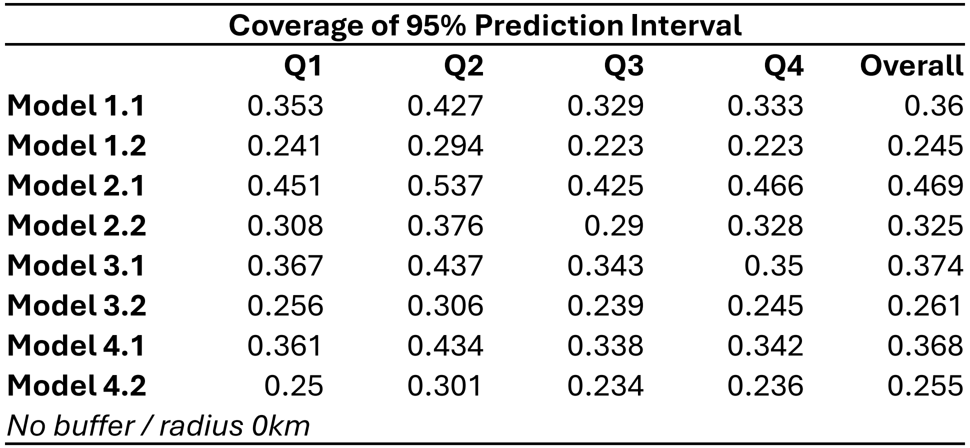
Radius 100km

Root Mean Square Error					
	Q1	**Q2**	**Q3**	**Q4**	**Overall**
Model 1.1	6.1	4.16	7.81	9.27	7.12
Model 1.2	6.1	4.16	7.81	9.27	7.12
Model 2.1	5.46	3.69	7.08	8.38	6.42
Model 2.2	5.46	3.69	7.08	8.38	6.42
Model 3.1	6.1	4.16	7.73	9.25	7.09
Model 3.2	6.1	4.16	7.66	9.24	7.07
Model 4.1	6.1	4.16	7.78	9.26	7.11
Model 4.2	6.1	4.16	7.74	9.25	7.1

Radius 200km

Coverage by Quarters

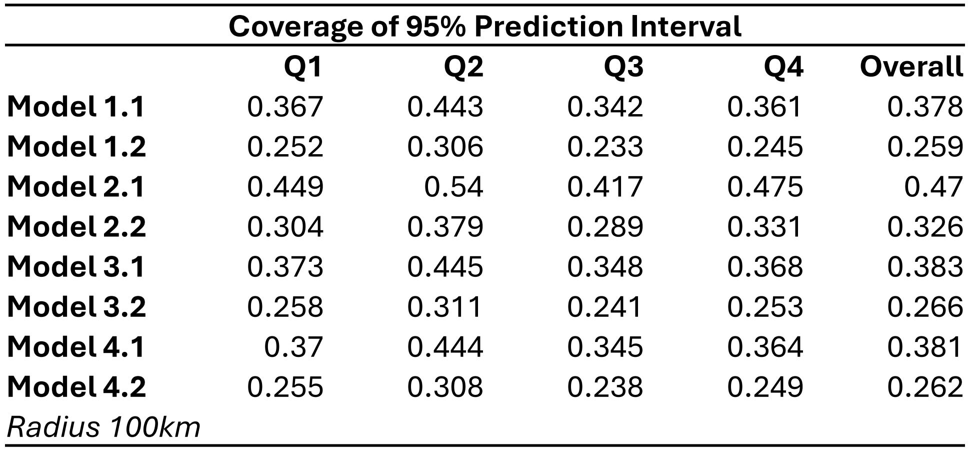
- Model 2.1 has the largest coverage probability overall and in all quarters.
- PM2.5 are predicted the most accurately in Q2 (Apr - Jun) by all models.



Coverage of 95% Prediction Interval

	Q1	**Q2**	**Q3**	**Q4**	**Overall**
Model 1.1	0.353	0.427	0.329	0.333	0.36
Model 1.2	0.241	0.294	0.223	0.223	0.245
Model 2.1	0.451	0.537	0.425	0.466	0.469
Model 2.2	0.308	0.376	0.29	0.328	0.325
Model 3.1	0.367	0.437	0.343	0.35	0.374
Model 3.2	0.256	0.306	0.239	0.245	0.261
Model 4.1	0.361	0.434	0.338	0.342	0.368
Model 4.2	0.25	0.301	0.234	0.236	0.255

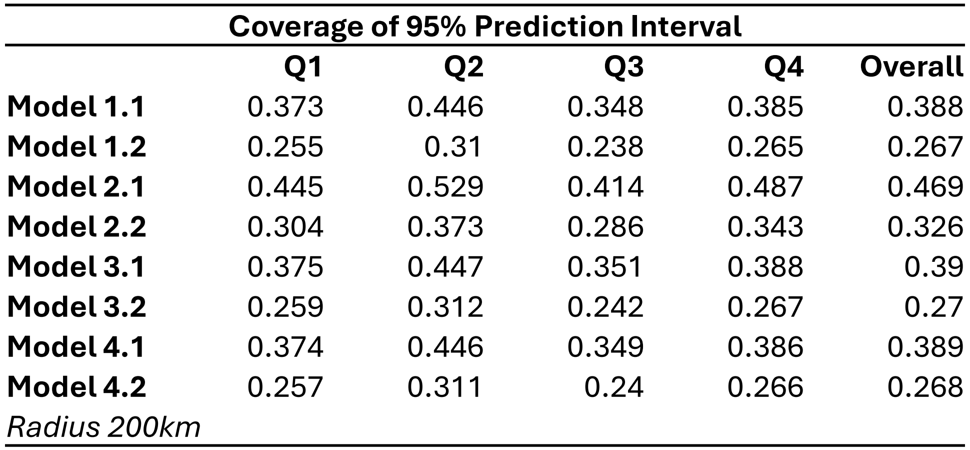
No buffer / radius 0km



Coverage of 95% Prediction Interval

	Q1	**Q2**	**Q3**	**Q4**	**Overall**
Model 1.1	0.367	0.443	0.342	0.361	0.378
Model 1.2	0.252	0.306	0.233	0.245	0.259
Model 2.1	0.449	0.54	0.417	0.475	0.47
Model 2.2	0.304	0.379	0.289	0.331	0.326
Model 3.1	0.373	0.445	0.348	0.368	0.383
Model 3.2	0.258	0.311	0.241	0.253	0.266
Model 4.1	0.37	0.444	0.345	0.364	0.381
Model 4.2	0.255	0.308	0.238	0.249	0.262

Radius 100km



Coverage of 95% Prediction Interval

	Q1	**Q2**	**Q3**	**Q4**	**Overall**
Model 1.1	0.373	0.446	0.348	0.385	0.388
Model 1.2	0.255	0.31	0.238	0.265	0.267
Model 2.1	0.445	0.529	0.414	0.487	0.469
Model 2.2	0.304	0.373	0.286	0.343	0.326
Model 3.1	0.375	0.447	0.351	0.388	0.39
Model 3.2	0.259	0.312	0.242	0.267	0.27
Model 4.1	0.374	0.446	0.349	0.386	0.389
Model 4.2	0.257	0.311	0.24	0.266	0.268

Radius 200km

Conclusion

- **Model 2.1** and **2.2** have the smallest RMSE values, and **Model 2.1** attains the largest coverage probability of the 95% prediction interval. Therefore, **Model 2.1** is identified as our final model:

$$y_i | \mu_i, \sigma_\epsilon \sim N(\mu_i, \sigma_\epsilon^2)$$
$$\mu_i = \beta_0(s_i) + \beta_1 X_i + u(s_i)$$
$$u(s_i) | \sigma, \rho \sim GP(0, \Sigma)$$

- $\beta_0(s_i)$ is the intercept as a spatially varying coefficient.
- β_1 is the coefficient of the exposure.
- $u(s_i)$ is a spatial random effect that follows a zero-mean Gaussian process with Matérn covariance function.

Model 2.1 Plots

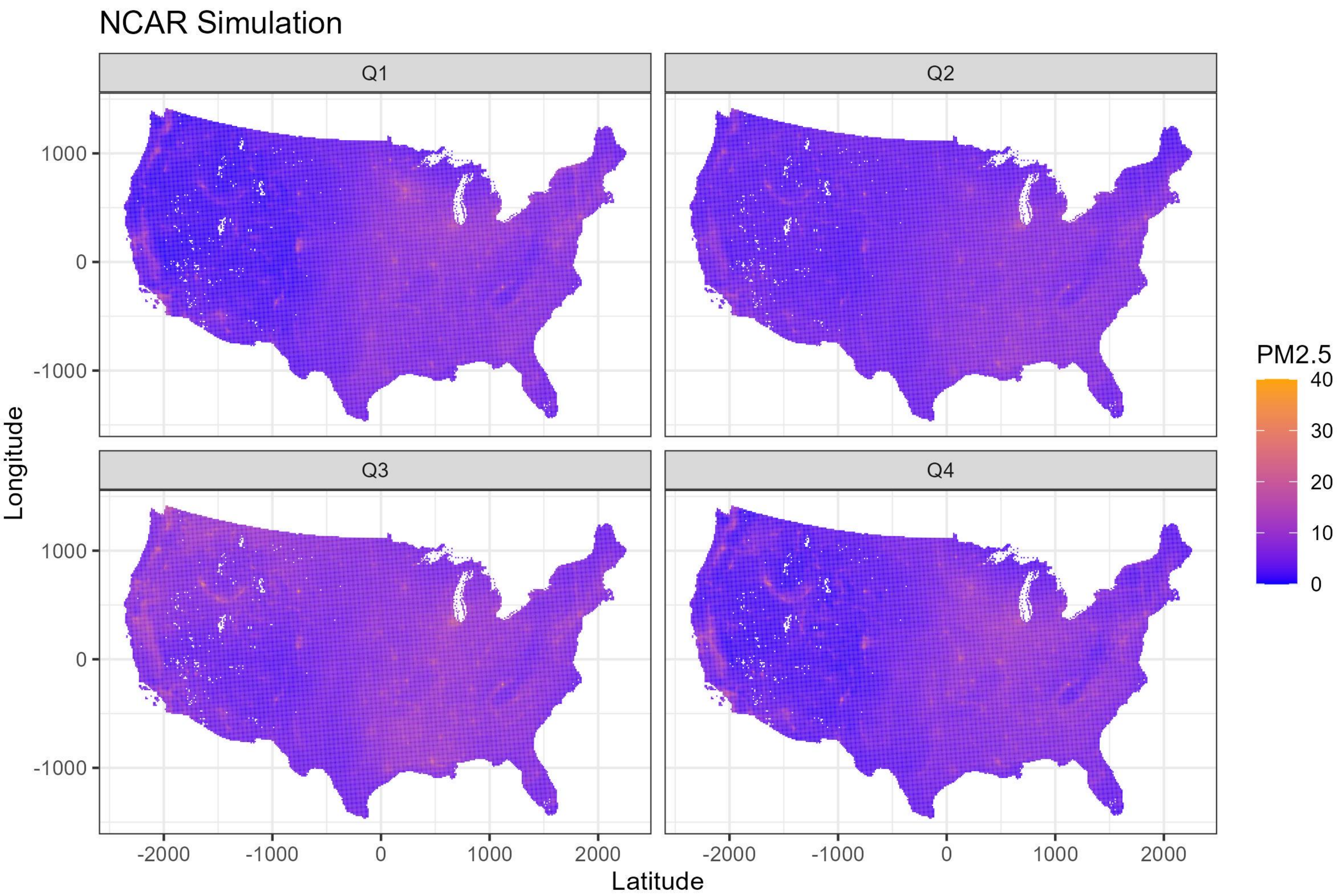
- PM2.5 measurements are predicted on the 12-by-12km grid using the chosen final model 2.1.

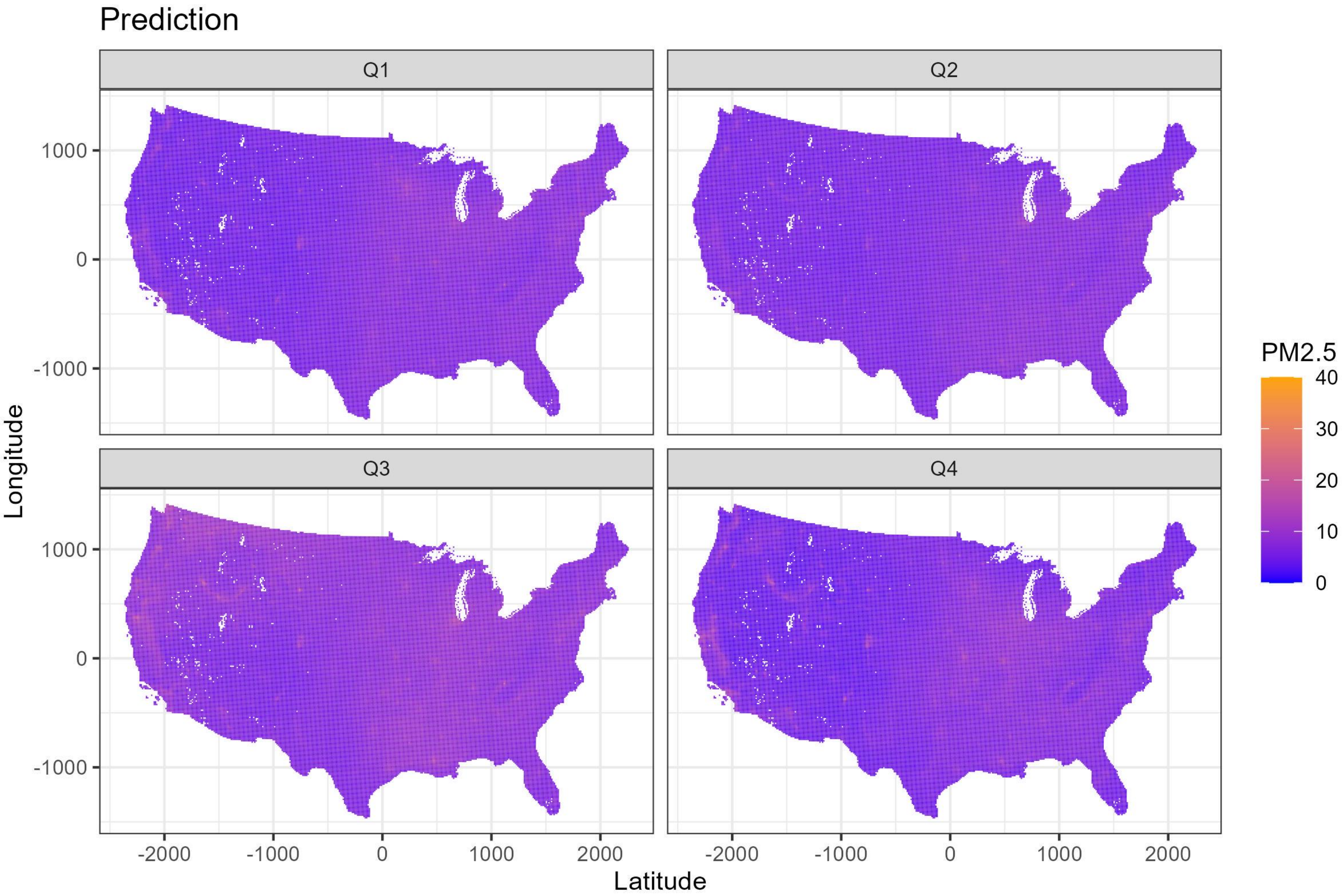
Values at all grid locations by quarter:

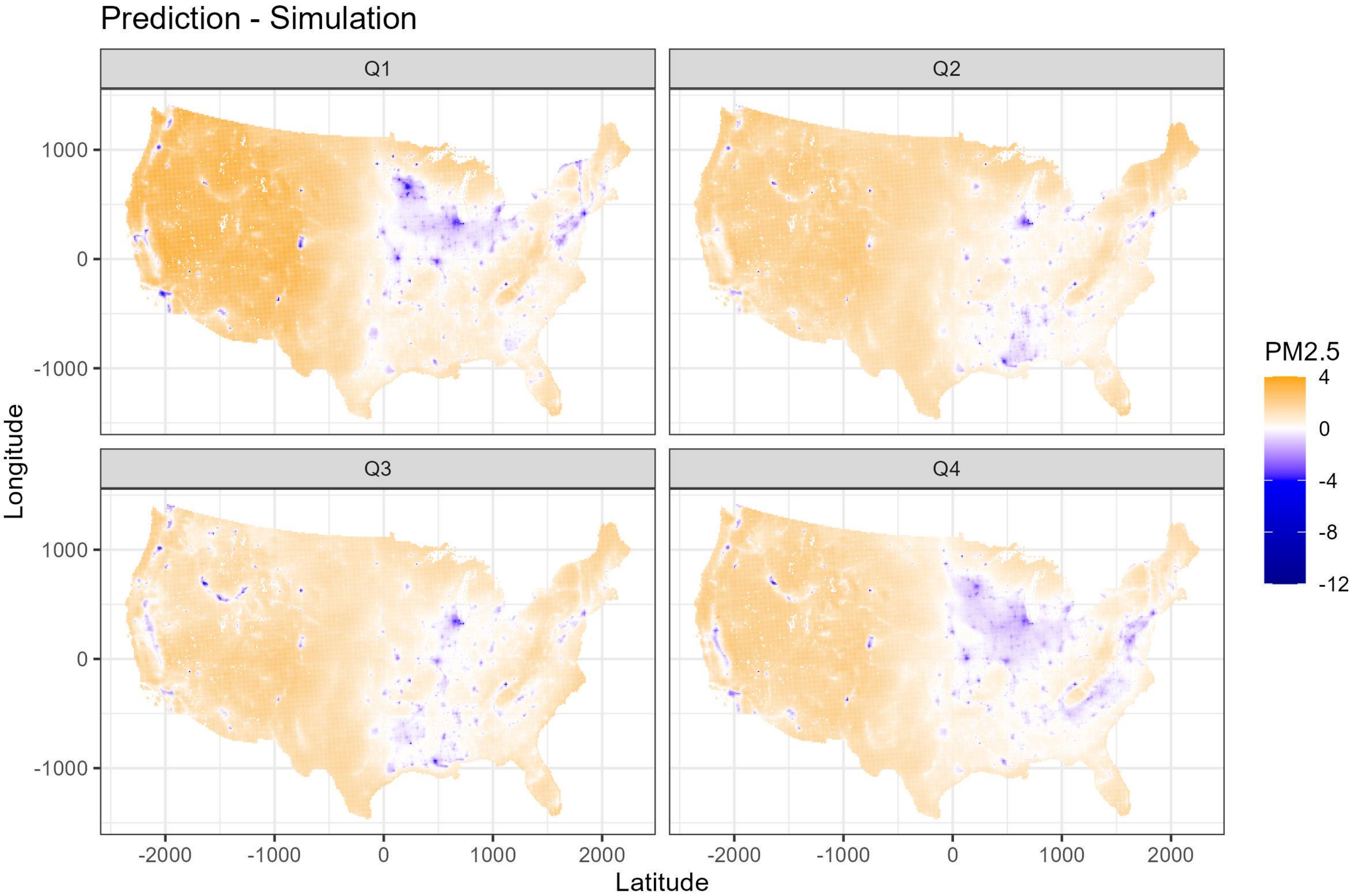
1. Mean PM2.5 NCAR simulation values (exposure)
2. Mean PM2.5 prediction values (outcome)
3. Mean differences between the prediction and the simulation values
4. Mean standard deviation of the predictions
5. Mean prediction interval lengths ($\approx 2 * 1.96 * \sqrt{(prediction\ sd)^2 + (Gaussian\ precision)^{-1}}$)

95% Error bars:

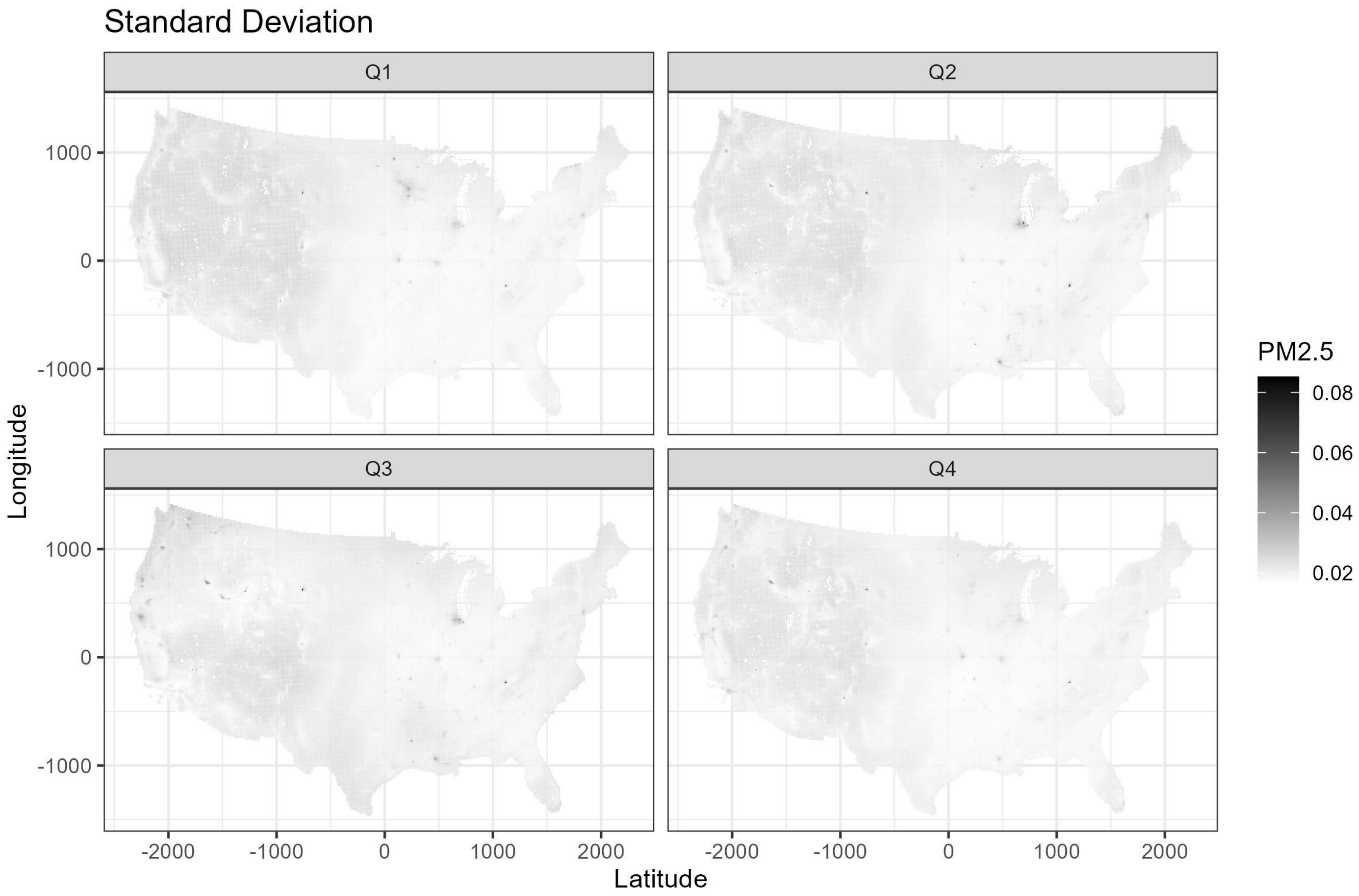
1. Intercept β_0 estimates vs. week
2. Covariate’s coefficient β_1 estimates vs. week
3. Range ρ of the spatial random effect estimates vs. week
4. Marginal standard deviation σ of the spatial random effect estimates vs. week

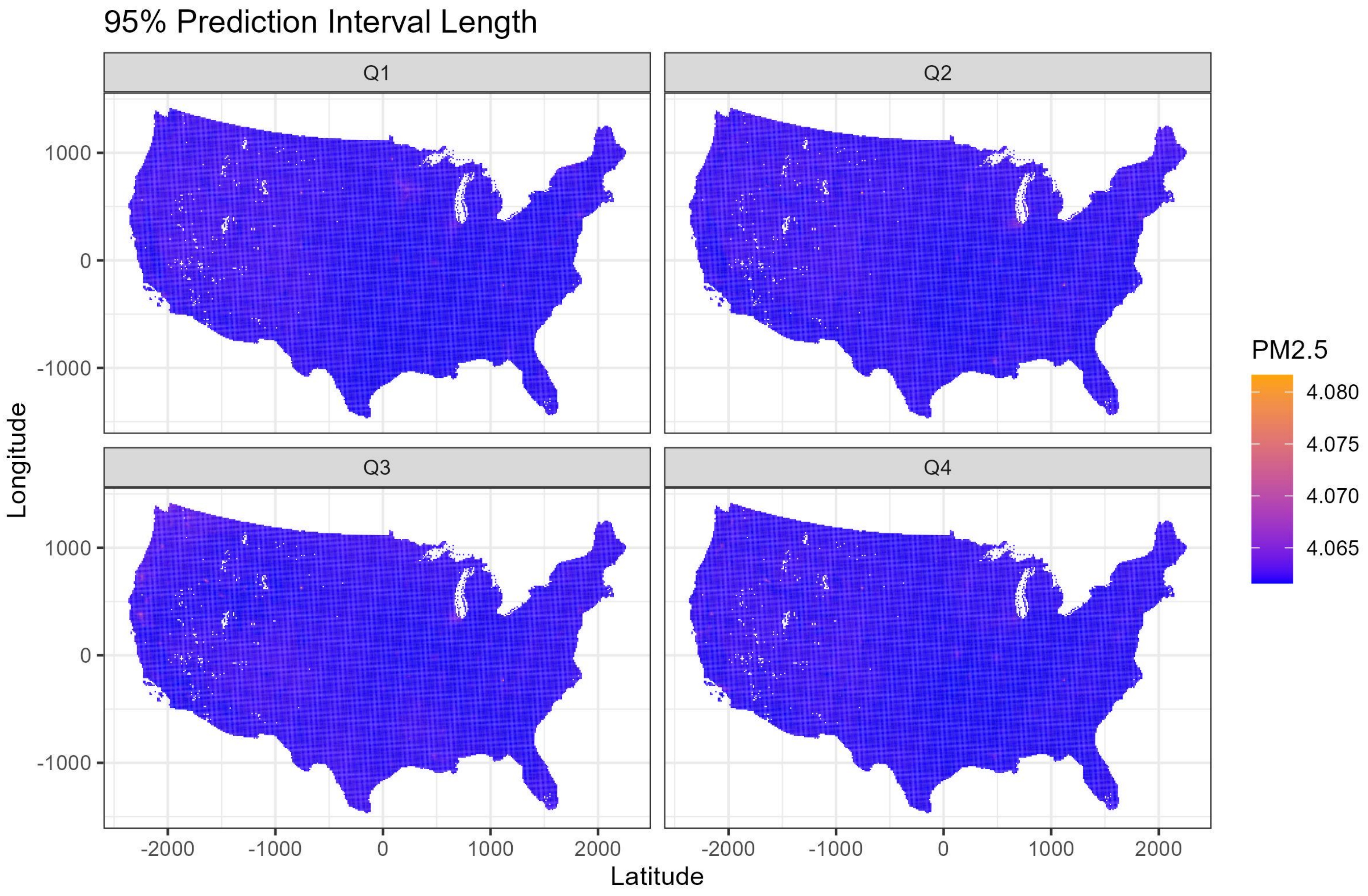

NCAR Simulation
Q1
Q2
Q3
Q4
1000
0
-1000
1000
0
-1000
Longitude
-2000
-1000
0
1000
2000
-2000
-1000
0
1000
2000
Latitude
PM2.5
40
30
20
10
0


Prediction
Q1
Q2
Q3
Q4
1000
0
-1000
Longitude
-2000
-1000
0
1000
2000
Latitude
PM2.5
40
30
20
10
0

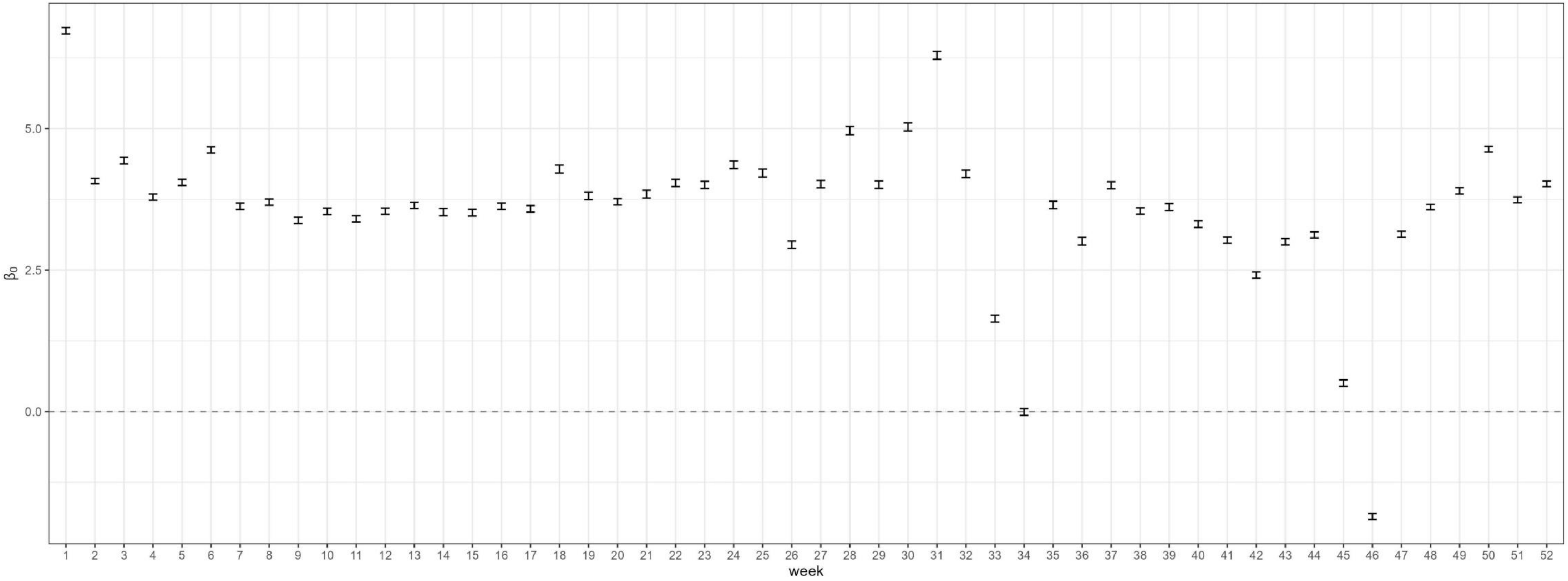


Note. PM2.5 are mean values by quarter.

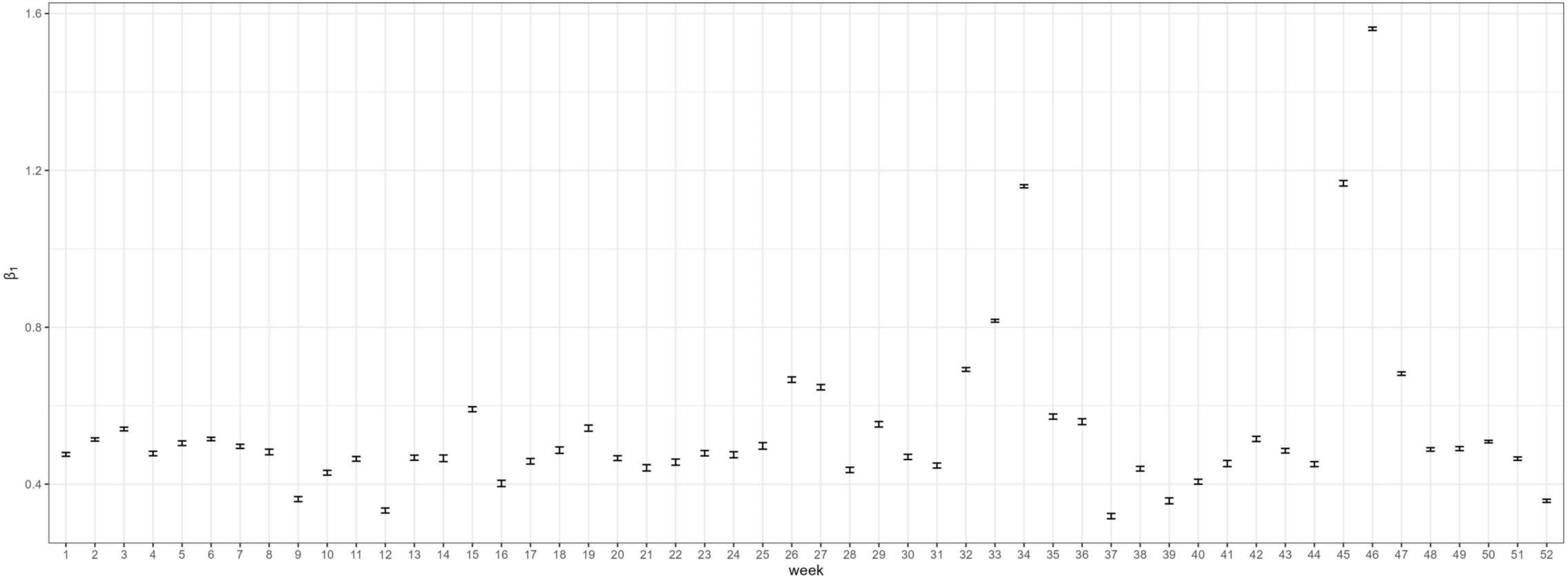

Standard Deviation
Q1
Q2
Q3
Q4
Longitude
1000
0
-1000
Latitude
-2000
-1000
0
1000
2000
PM2.5
0.08
0.06
0.04
0.02


95% Prediction Interval Length
Q1
Q2
Q3
Q4
Longitude
Latitude
1000
0
-1000
-2000
-1000
0
1000
2000
PM2.5
4.080
4.075
4.070
4.065

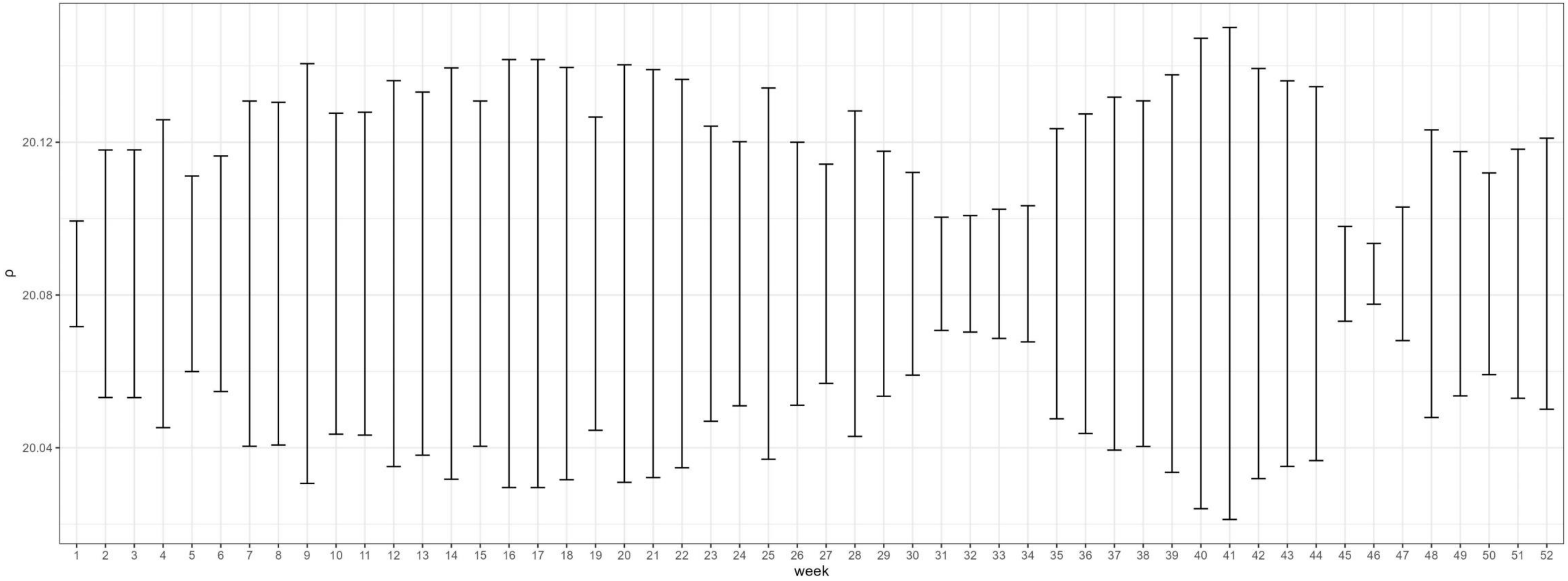
Intercept β_0 estimates vs. week



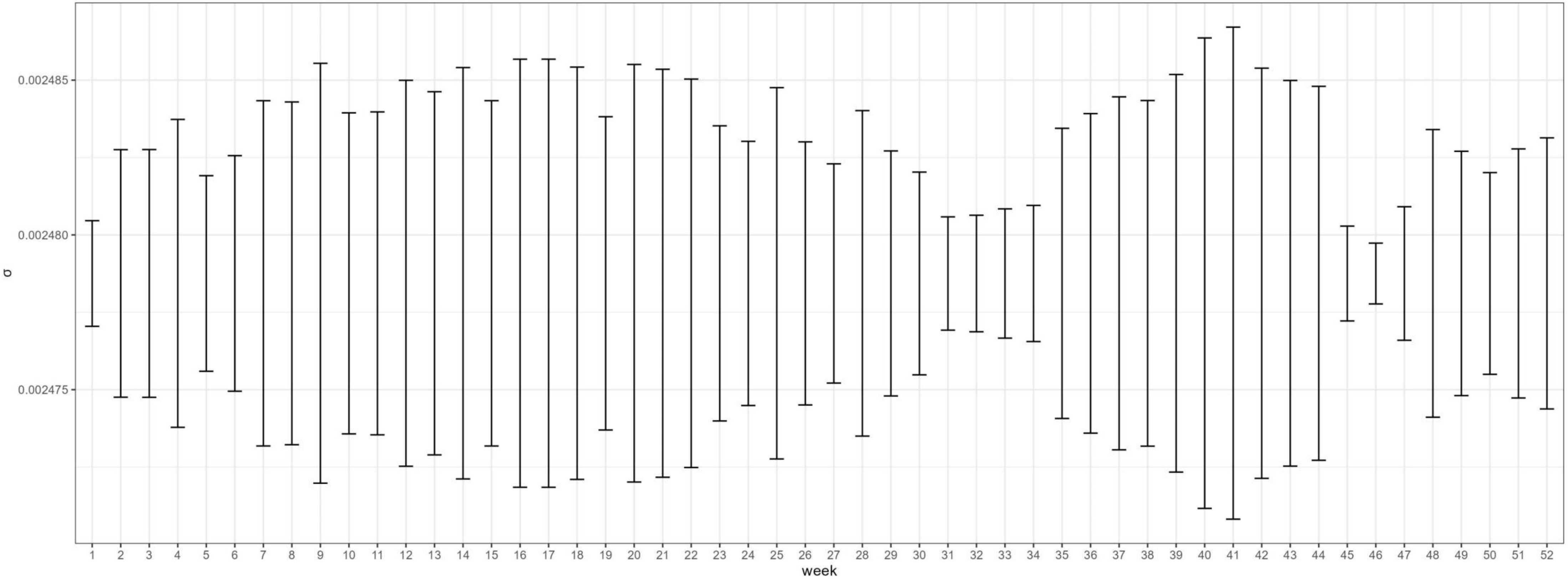
Covariate’s coefficient β_1 estimates vs. week



Range ρ of the spatial random effect estimates vs. week



Marginal standard deviation σ of the spatial random effect estimates vs. week


σ
week