Spatio-temporal Bayes Regression with INLA in Statistical Downscaling Modeling for Estimating West Java Rainfall

Ro'fah Nur Rachmawati[1,2], Anik Djuraidah[1*], Aji Hamim Wigena[1], I Wayan Mangku[3]
{[*)]Corresponding author: anikdjuraidah@apps.ipb.ac.id}

Statistics Department, IPB University, Bogor, Indonesia 16680[1]
Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480[2]
Mathematics Department, IPB University, Bogor, Indonesia 16680[3]

Abstract. Statistical downscaling (SD) is one of the techniques used in climate modeling by utilizing global scale data using general circulation models (GCM) output data, to obtain conclusions on a local scale such as rainfall. Currently, the inference of Bayes spatio-temporal regression in SD modeling is still used MCMC method, with convergence issue problem and very high demands for computational resources. When the spatio-temporal model is complex and designed hierarchically, MCMC computing becomes inefficient. Therefore, this paper aims to predict observed and unobserved locations, using Bayes spatio-temporal model with efficient, fast, accurate and developed inference method, INLA. The response variable is monthly rainfall at 57 locations in West Java, Indonesia, observed from 1981-2017 and assumed to have normal distribution. The explanatory variables consist of spatial and temporal random effects and fixed effects of monthly precipitation GCM with 8x5 dimensions (40 variables) and the dimension is reduced with PCA. Our model successfully predicts monthly rainfall for observed and unobserved locations using spatial characteristics from nearly locations, and primely capture the monthly rainfall trends in annually cyclic behavior. The correlations between predict and real rainfall data is about 0.8 (for 0.65, 0.8 quantile) and 0.7 (for 0.95, 0.975 high quantile) with RMSEP is 151 for low (0.65) quantile. At the end of the research results, we present the regional rainfall for the entire West Java region. The eastern part near the central Java border has higher rainfall, as well as the west, while the north and south have lower rainfall.

Keywords: Bayes spatio-temporal, INLA (integrated nested Laplace approximation), PCA (principal component analysis), statistical downscaling, West Java rainfall region.

1 Introduction

Statistical downscaling (SD) is a statistical technique used to conduct future projections from responses in the form of local climate data, with explanatory variables using global circulation model (GCM) output data. Several recent studies in various countries actively used GCM for SD modeling including [12] who studied present and future climate projections in China and [2] used SD modeling to achieve projections of precipitation extremes in New England. In GCM data, global climate variables are simulated on each grid for each atmosphere layer, therefore GCM are in the form of grids, rough spatial resolution, large

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290346

dimensions and multicollinearity. There are several studies for weather and climate prediction by utilizing various techniques to overcome these data problems. [7] used the group lasso regularization technique, [11] using lasso and principal component analysis, [10] using functional principal component, and [6] using elastic net regularization technique. These studies use GCM data to predict the amount of rainfall in several regions of Indramayu Regency, West Java, Indonesia.

Climate modeling is a representation of complex phenomena, which involve spatial, temporal, spatial and temporal interactions, regional topography and other influences. The Bayes method is one solution in representing these complex phenomena, because the complexity of the model can be represented by designing a hierarchical structure for data and its parameters. Several studies have focused on Bayes modeling with MCMC inference for spatio-temporal data, including [13] analyzed annual minimum temperatures for the past 6 decades in Mindland, China, and assumed the data to have GEV distribution, data is assumed to be stationary and arranged hierarchically. [4] used a dynamic linear model based on GEV distribution on monthly maximum wind speed data.

MCMC methods are extensively used for Bayesian inference, however their limitation resides in their computational burden. Moreover, MCMC method in spatio-temporal modelling is no longer efficient and may caused non-convergence to the posterior distribution [1]. To overcome this issue, the integrated nested Laplace approximations (INLA), a deterministic algorithm proposed by [8], has proven capable of providing accurate and fast results [8], [9]. [5] used the hierarchical Bayes method with INLA inference to model daily precipitation data in Norway.

Although INLA has been applied to spatio-temporal modeling for weather and climate prediction, the application of INLA inference to SD modeling is still rare or even not yet available. Therefore, the uniqueness of this paper is to combine SD modeling by involving GCM data as a global explanatory variable and spatio-temporal modeling involving local responses. Large dimension and multicollinearity of GCM data is overcome by using the principal component analysis (PCA), then the selected principal components are used as the explanatory variable in the spatio-temporal model whose inferences are obtained by INLA. In more detail, using the Bayes spatio-temporal in SD model, this paper aims to predict the local response of the observed and unobserved locations by assuming the response has a normal distribution. The model is based on linear additive mixed models, where in addition to the GCM explanatory variables as a fixed component, there are also spatial and temporal random components that can flexibly capture spatial and temporal dependencies. In the remainder of this paper, we present the datasets description, the detailed methodology of Bayes spatio-temporal in SD model with INLA inference, reported results and discussions. Some concluding remarks and possible future development are also summarized in the end of this paper.

2 Materials: Data Sources and Descriptions

2.1 Rainfall as a Local Response and GCM as a Global Fixed Explanatory Variables

Rainfall as a Local Response. The local response dataset, derived from Indonesian National Board for Disaster Management (or simply BMKG), consisted of monthly rainfall accumulations recorded in milliliters at 57 stations during the period 1981 – 2017. The data

were divided into training set (January 1981 – December 2005) which was made available to spatio-temporal model, and a validation set (January 2006 – December 2017) which was used to assess quantile predictions. In the training period, we have a mixed dataset comprised of 12 unobserved (poor station) (station 4, 6, 13, 22, 26, 30, 32, 33, 40, 42, 47 and 50) and 45 observed (rich station) (the rest stations). In validation set, the data varies greatly. Only 1 station that have full sample size $n = 144$, while as many 8 stations that have sample sizes $n < 30$ with minimum sample size $n = 2$. The exact coordinates of stations are shown in Figure 1, stations location is the location of rainfall observation in West Java, Indonesia.

GCM as a Global Fixed Explanatory Variables. The GCM data used was precipitation, which was issued by the National Centers for Environmental Prediction (NCEP) in the form of Climate Forecast System Reanalysis (CFSR). In this study the variables used were the average precipitation rate taken from the website https://rda.ucar.edu/. The GCM data has a grid type and covers all of the West Java, with grid size is 2.5° × 2.5° and 8x5 dimensions (40 variables).

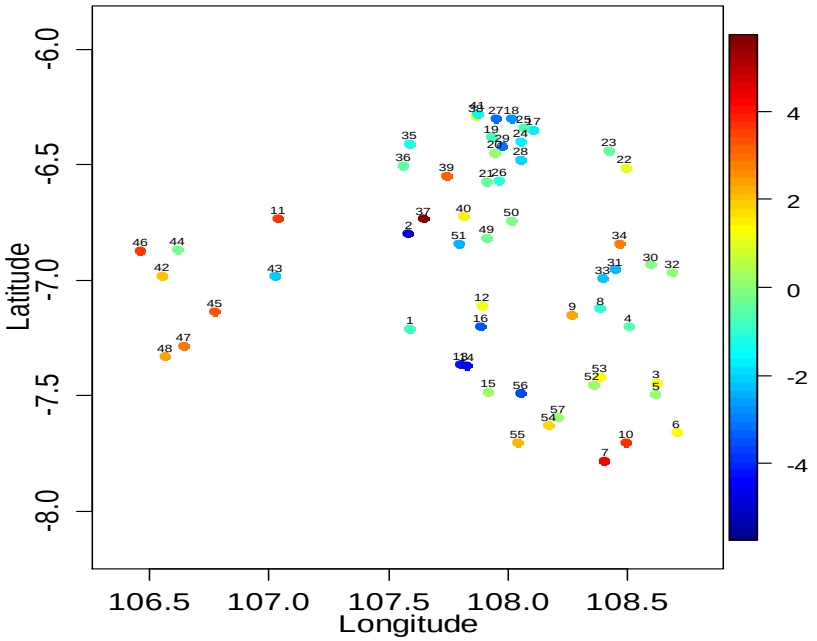


Fig. 1. Map of monitoring locations colored according to the estimated spatial random effect in equation (1).

3Methods

The PCA is used as a pre-processing technique to obtain latent variables that are orthogonal and are linear combinations of the covariates. In this study used a cumulative variance proportion ≥ 90% and the eigen value > 1. The scree-plot not shown in this paper,

but it suggests to take the number of latent orthogonal variables as many as 3 PCs. These first three PCs then will used in our proposed Bayes spatio-temporal in SD model (equation (1)).

3.1 Spatio-temporal Bayes in SD Modeling and INLA Inference

Spatio-temporal Bayes in SD Modeling. The proposed model based on linear additive mixed property that have high flexibility and interpretability by involving the spatial and temporal random effects as explanatory variables. Let$Y(s,t)$is monthly rainfallin specific spatial location $s = \{s_1, s_2, \dots, s_{57}\} \in S$where$S \in \mathbb{R}^2$and time$t \in \{1,2,\dots,n\}$. $Y(s,t)$is assumed to have normal distributionwith mean $\eta(s,t) > 0$and varianceσ_e^2. To represent location and time diversity in spatio-temporal parameters, formulated regression equations are inhaled additives, which are summations of spatial and temporal random components which are assumed to be separable, as follows:

$$\eta(s,t) = \beta_0^{\text{Nor}} + x^{\text{Nor}}(s) + x^{\text{Nor}}(t) + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3, \quad (1)$$

where β_0^{Nor} is the intercept and $\beta_i, i = 1,2,3$are coefficients for the selected PCs of GCM data as fixed effects, assumed $\beta_1, \beta_2, \beta_3, \beta_0^{\text{Nor}} \sim \text{Normal}(0, 10^5), x^{\text{Nor}}(s)$ and $x^{\text{Nor}}(t)$ are the spatial and temporal random effects respectively. We assume $\log \tau = 1/\sigma_e^2 \sim \log \text{Gamma}(1, 0.00005)$.

The spatial random effects are the realization of Gaussian process and the spatial dependency follow Matérn covariance function as

$$\text{Cov}\{x(s_1), x(s_2)\} = {\tau_s}^{-1} \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}h}{\psi}\right)^v K_v \left(\frac{\sqrt{2v}h}{\psi}\right), \quad (2)$$

with$\sigma^2 = {\tau_s}^{-1} > 0$is the variance of locations$s$, $h = \|s_1 - s_2\|$is the Euclidean distance between locations$s_1, s_2 \in \boldsymbol{S}$, $K_v$is the modified Bessel function of order v, $v = 1$and$\psi > 0$ is a range parameter that plays animportant role in the process of spatial smoothing. We perform the distance so that the correlation near 0.1 using local regression method and derived ψ is about 106 km.

Random walk (RW) of order 2 generally used to model response functions and is a key role in the temporal smoothing process. RW of order 2 with density function as follows:

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\sum_{i=2}^{n-1}(x_{i-1} - 2x_i + x_{i+1})^2\right), \quad (3)$$

where$\boldsymbol{x} = [x_1, x_2, \dots, x_n]^{\boldsymbol{T}}$. The temporal effects $x^{\text{Nor}}(t)$ are defined assumed to have a normal distribution that follows a random walk of order 2 with mean 0 and precision $\tau_t > 0$ is the temporal smoothing parameter, assumed in monthly basis and annually cyclic.

INLAInference. Let$y(s_i, t_i) = (y_1, y_2, \dots, y_m) = \boldsymbol{y}$, $i = 1,2,\dots,m$ is the observation data with the latent Gauss explanatory variable declared as $\boldsymbol{\eta} = (\eta_1, \eta_2, .., \eta_m)^T$then$\eta_i = \beta_0 + x(s_i) + x(t_i)$, $\boldsymbol{\theta}_y$ is a vector for hyperparameters for data distribution y, and vector for hyperparameters for spatial and temporal random component is$\boldsymbol{\theta}_x$. So that the distribution of prior hyperparameters is defined as$\pi(\boldsymbol{\theta})$with$\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_x)$, with Gaussianprobability$\boldsymbol{x}$can be written as$\pi(\boldsymbol{x}|\boldsymbol{\theta}_x)$. Let$\pi(y_i|\eta_i, \boldsymbol{\theta}_y)$is a*likelihood*from$y_i$with condition the explanatory variables$\eta_i$and*likelihood*from hyperparameters$\boldsymbol{\theta}_y$, then the joint posterior for vector $\boldsymbol{x}$and hyperparameters$\boldsymbol{\theta}$can be determined with Bayesian formulas as follows

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{x}^T Q(\boldsymbol{\theta}_x)\boldsymbol{x} + \sum_{\boldsymbol{i}} \log \pi(y_i|x_i, \boldsymbol{\theta}_y) + \log \pi(\boldsymbol{\theta}) + \log K(\boldsymbol{\theta}_x)\right\} \quad (4)$$

where$Q(\boldsymbol{\theta}_x)$is a precision matrix (i.e inverse of the covariance matrix), and$K(\boldsymbol{\theta}_x)$is a normalized constant derived from hyperparameters$\boldsymbol{\theta}_x$.

INLA is an analytical Bayes-based inference, which can be applied to the generalized additive model that is complex and hierarchical, with the latent Gauss variable as in the equation (4). INLA produces an approximation to the two posterior distributions of the following single variables:

$$\pi(\theta_k|\boldsymbol{y}) = \int \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})\mathrm{dx}\,\mathrm{d}\boldsymbol{\theta}_{-k}, \tag{5}$$

$$\pi(x_i|\boldsymbol{y}) = \int\int \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}_{-i}\mathrm{d}\boldsymbol{\theta} = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\,\pi(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}, \tag{6}$$

where$i = 1,2, \ldots, \tilde{m}$with$\tilde{m}$ is the length of the vector $\boldsymbol{x}$. The Laplace approximation is applied nestedly, the first approximation is done to determine the posterior distribution of the hyperparameter $\pi(\boldsymbol{\theta}|\boldsymbol{y})$at equation (5), and the second approximation is done to the posterior distribution of parameters $\pi(x_i|\boldsymbol{y})$atequation (6). For more details on INLA estimation procedure and its statistical properties can be seen in [3], [9].We perform the model inference using the R package R-INLA, which is convenient, easy to use and available free in a highly developed website for Bayesian inference with INLA, http://www.r-inla.org/.

4Results and Discussions

Results and Discussions. Here, we provide further results, interpretations and discussions for Bayes spatio-temporal in SD model as in equation (1). We derived spatial and temporal random component as in Figure 2. Black dots and curves denote posterior mean and blue segments and curves are 95% pointwise credibility interval. The spatial random components have significant and almost significant for observed locations. As expected, the credible interval for poor or unobserved location is quite large compared to those at the rich stations, except for station 13, it is because, station 13 is very close to the rich station 14, which confirm that the proposed model successfully borrows the strength of rich stations to nearly locations. The captured temporal trend of monthly rainfall is clearly visible and very significantly different from zero, so it provides very important information for the behavior of the annual rainfall cycle. It has significant positive effects in rainy season at the end of the year, increase from October to January, and decrease to April. The dry season occurs until the beginning of October with significant negative effects.

The estimated spatial random effects from model in equation (1) is presented in Figure 1. The poor stations have unique spatial effects and the model shows that nearly locations have the same spatial characteristics. Table 1 presents a summary of estimated posterior of β_0^{Nor}, the regression intercept, and $\beta_i, i = 1, 2, 3$, the coefficients for the selected PCs. Overall, the posterior mean of fixed effects has significant effects, except for β_2. The regression intercept $\beta_0^{\mathrm{Nor}} = 12.39$ means that the monthly rainfall is not center at zero, while β_1 and β_3 have significant effects, which means that the selected GCM variables have a significant influence on monthly rainfall in each spatial location.

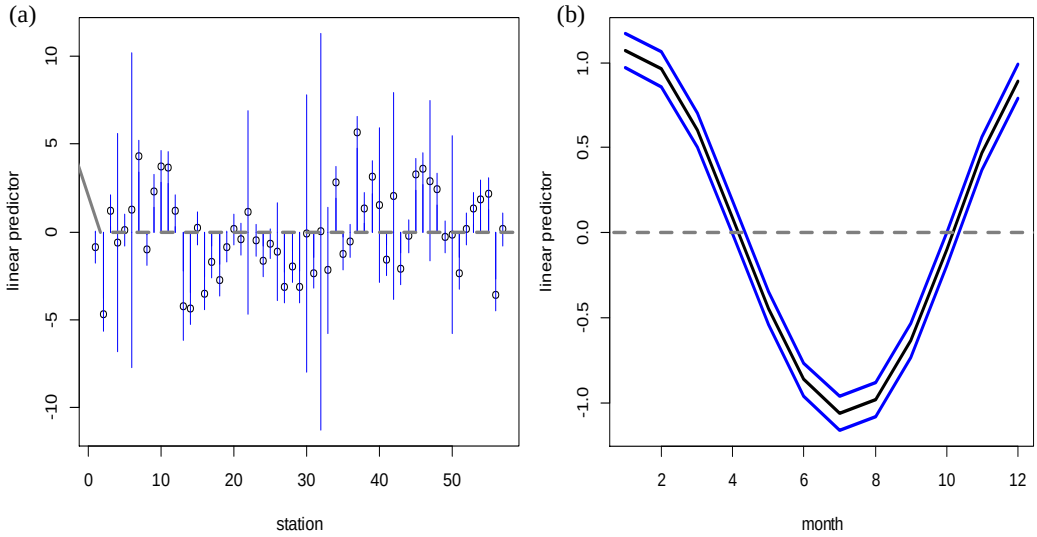


Fig. 2. Spatial random effects (a) and temporal (monthly) (b) random effects.

Table 1. Estimated posterior of fixed effects.

Fixed effects	Mean	Standard Deviation	Credibility Interval
β_0^{Nor}	12.39	0.29	(11.87, 12.97)
β_1	0.52	0.03	(0.47, 0.58)
β_2	-0.01	0.05	(-0.10, 0.09)
β_3	-0.66	0.06	(-0.77, -0.54)

Our goal is to predict quantile monthly rainfall for observed and unobserved locations from derived model parameters. Using validation data as stated in previous section, the results are presented in Table 2. The higher the quantile value, the smaller the correlation mean and for the opposite for the RMSEP. The estimated rainfall has a highest correlation and the lowest RMSEP for low quantile value (quantile 0.65), this show that the assumption of a normal rainfall distribution produces good results only for low quantile value.

We also do an exploration to find out more details about the RMSEP values for rich and poor stations status along with terrain types such as lowland with 0-200 meters above sea level, medium land (> 200-500) and highland (> 500). The result is presented in Table 3. For low quantile (quantile 0.65) and medium quantile (quantile 0.8), there is a considerable difference between RMSEP of rich and poor station. Rich station with a sufficient number of observations having a much smaller RMSEP value compared to poor stations. For the types of terrain, the lowest RMSEP is lowland, this result is related to the variance of rainfall from each type of terrain. From local rainfall data, the medium land has the highest variance of rainfall, followed by the highlands, and the lowlands have the lowest variance of rainfall.

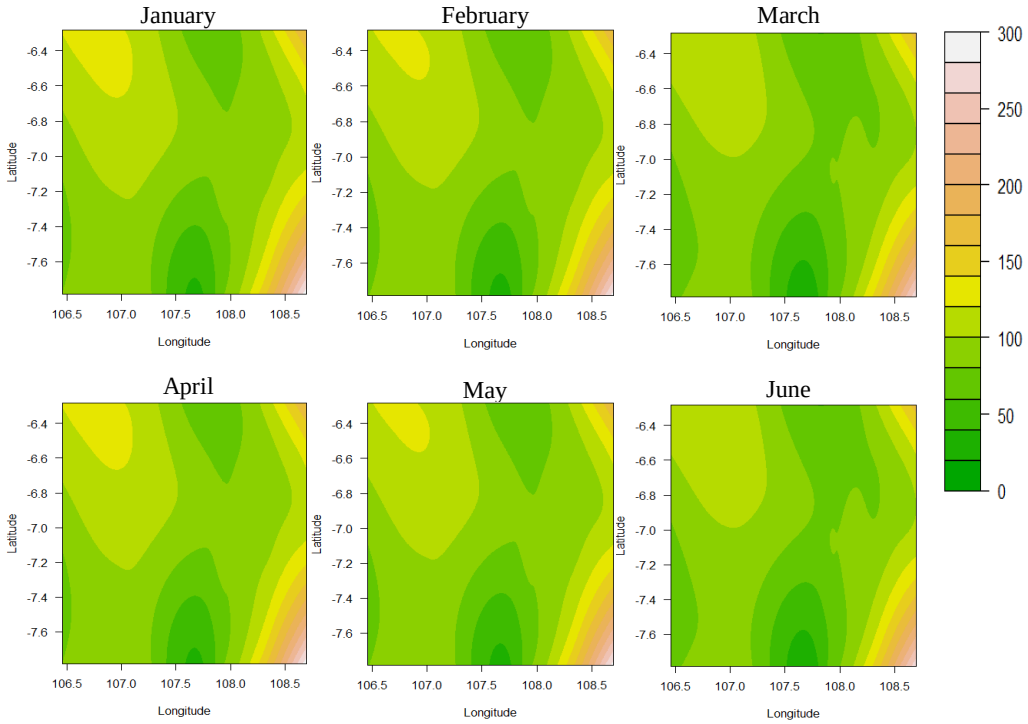
Table 2. Goodness of estimated quantile monthly rainfall

Quantile	CorrelationMean	RMSEP Mean
0.65	0.81	150.66
0.80	0.80	199.46
0.95	0.74	293.76
0.975	0.71	316.33

Table 3.RMSEP mean of various stations.

Quantile	Rich	Poor	Terrain types		
			Low	Medium	High
0.65	148.32	205.87	119.02	226.64	166.06
0.80	195.35	236.41	161.40	287.17	219.83
0.95	286.15	284.82	243.79	433.00	308.48
0.975	308.66	293.81	262.65	469.98	330.10

At the end of the results of this research, from the estimated model values in equation (1), we present rainfall zoning for all regions in West Java province, Indonesia, using the classical non- parametric method i.e. local regression. Using 0.65 quantile, Figure 3 presents the regional rainfall of the entire West Java Province. The eastern part near the central Java border has higher rainfall, as well as the west which borders DKI Jakarta and Banten, while the north and south have lower rainfall.



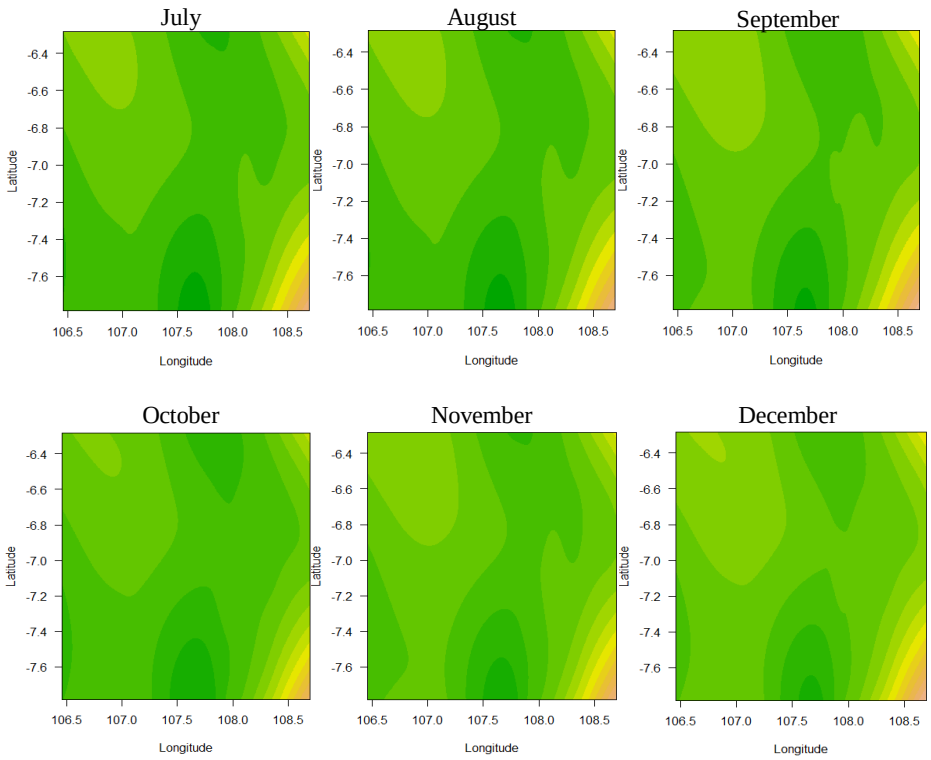


Fig. 3. Regional rainfall of West Java, quantile 0.65.

4Conclusions and Possibilities for Future Research

This paper predicts the amount of monthly rainfall even for locations where no data have been recorded. The proposed model use Bayesian approach with linear mixed in SD modeling, with additively involve spatial and temporal random component to capture the flexibility in hierarchical spatio-temporal data modeling, and fixed effects derived from selected GCM precipitation data with dimension reduction and multicollinearity data handling using the PCA. The spatial characteristic and annually cyclic temporal trend are derived with significant or almost significant for rich stations and a few for poor stations. It means the proposed model systematically able to capture spatial and temporal linkages between locations. Generally, the selected GCM precipitation data has a significant influence on local rainfall conditions, this indicates that the selected GCM variables can also be used for projection of long-term climate conditions for the West Java region. The proposed model successfully predicts even for the unobserved locations with good correlation mean overall, with RMSEP mean is about 151 millimeters for low quantile 0.65.

For spatio-temporal cases with many unobserved locations and imperfect validation data, our predicted quantile is good overall although certain aspects could be improved. The Bayes spatio-temporal in SD model could be made more complex if required by the context. We can enhance other additional fixed effects like altitude or by including more complex random

effects like space-time interaction. However, the proposed model restricted to the above structure, which was flexible enough and provided robust and interpretable results. Dimension reduction results are very dependent on the reduction technique used. Other dimensional reductions can be used such as lasso [11] or functional principal component because GCM data is believed to form a function [10] The resulting regional rainfall of West Java will depend on the location of the selected sample points and the non-parametric methods used. West Java's topography is very diverse, consisting of low and high land, hilly natural conditions and bordering the oceans, making meteorological conditions very varied. This means that the processes involved in this area are processes that are not stationary. Estimating the non-stationary data is a difficult task, however this work has positively contributed in advancing methods in the current literature.

Acknowledgments. This work is fully supported by Kemenristek DIKTI (KementrianRisetTeknologi dan Pendidikan Tinggi) of Indonesia.

References

[1] Blangiardo, M., and Cameletti, M.:Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons(2015)

[2] Laflamme, E. M., Linder, E., and Pan, Y.: Statistical downscaling of regional climate model output to achieve projections of precipitation extremes. Weather Climate Extremes.12:pp. 15-23(2016)

[3] Lindgren, F., & Rue, H.: Bayesian spatial modelling with R-INLA. Journal of Statistical Software. 63(2015).

[4] Mahmoudian, B., and Mohammadzadeh, M.: A spatio-temporal dynamic regression model for extreme wind speeds. Extremes.17(2): pp. 221-245(2014)

[5] Opitz, T., Huser, R., Bakka, H., and Rue, H.: INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. Extremes.pp. 1-22(2018)

[6] Permatasari, S. M., Djuraidah, A., and Soleh, A. M.:Statistical Downscaling with Gamma Distribution and Elastic Net Regularization. 2nd International Conference On Applied Statistics.(2017)

[7] Rachmawati, R. N., Pusponegoro N. H., Muslim A, Anwar K. N., and Sartono B.: Group LASSO for Rainfall Data Modeling in Indramayu District, West Java, Indonesia. 2nd International Conference on Computer Science and Computational Intellegence (ICCSCI). 1: pp. 190-197(2017).

[8] Rue, H., Martino, S., and Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B.71(2):pp. 319-392(2009)

[9] Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., Lindgren, F. K.: Bayesian computing with INLA: a review. Annual Review of Statistics.4:pp. 395-421(2017)

[10] Sari, W. J., Wigena, A. H., and Djuraidah, A.: Quantile Regression with Functional Principal Component in Statistical Downscaling to Predict Extreme Rainfall. International Journal of Ecological Economics Statistics.38(1):pp. 1-9(2017)

[11] Soleh A. M., Wigena A. H., Djuraidah A, and Saefuddin A.: Statistical Downscaling to Predict Monthly Rainfall Using Linear Regression with L1 Regularization (LASSO). Applied Mathematical Sciences. 9: pp. 5361 – 5369(2015)

[12] Tang J, Niu X, Wang S, Gaol H, Wang X, and Wu J.: Statistical downscaling and dynamical downscaling of regional climate in China: Present climate evaluations and future climate projections. Journal of Geophysical Research: Atmospheres. 121: pp. 2110–2129(2016)

[13] Yang, C., Xu, J., and Li, Y.: Bayesian geoadditive modelling of climate extremes with nonparametric spatially varying temporal effects. International Journal of Climatology: A Journal of the Royal Meteorological Society.36(12):pp. 3975-3987(2016)