ICIC Express Letters
Volume **14**, Number **1**, January **2020**

ICIC International ©2020 ISSN 1881-803X
pp. **89–96**

EXTREME DATA ANALYSIS USING GENERALIZED BAYES SPATIO-TEMPORAL MODEL WITH INLA FOR EXTREME RAINFALL PREDICTION

Ro'fah Nur Rachmawati[1,2], Anik Djuraidah[1,*], Aji Hamim Wigena[1] and I Wayan Mangku[3]

[1]Statistics Department
[3]Mathematics Department
Faculty of Mathematics and Natural Sciences
IPB University
Jl. Raya Dramaga, Kampus IPB Dramaga Bogor, West Java 16680, Indonesia
[*]Corresponding author: anikdjuraidah@apps.ipb.ac.id

[2]Statistics Department
School of Computer Science
Bina Nusantara University
Jakarta 11480, Indonesia
rofah.nr@binus.ac.id

Received July 2019; accepted October 2019

Abstract. *The Bayes spatio-temporal modeling has the advantage of estimating unobserved areas using similarities in spatial characters from adjacent locations and seeing changes in regional characteristics over time. Estimates for Bayes spatio-temporal modeling are still using the MCMC algorithm which requires a long computational time and becomes inappropriate to use if the model is hierarchically compiled and involves many parameters. This paper aims to predict the unobserved locations with fast, accurate and current developed estimation that is INLA. The modeling algorithms are divided into 3 main steps, Gamma distribution for overall data, Bernoulli for extreme data identification and Generalized Pareto for extreme data. To produce accurate predictive values, we innovatively purpose improvisations in determining spatial and temporal smoothing parameters, as well as determining the extreme value threshold using Measure of Surprise method. The spatio-temporal data is the monthly rainfall of 57 locations from West Java, Indonesia, observed from 1981-2017. The model produces satisfactory results: the spatio-temporal modeling improves the estimation of rainfall in many missing and completely unobserved data with the correlation between predictive and validation values about from 0.8 and RMSEP 137-195 mm for average to extreme rainfall, and 0.7 with RMSEP 224-229 mm for high extreme rainfall.*

Keywords: Extreme data analysis, Big data analysis, Bayes hierarchical spatio-temporal model, INLA

1. **Introduction.** Astronomically, a country with a tropical climate has a large variety of rainfall. This can lead to an increase or decrease in extreme rainfall which has the potential to cause hydrometeorological disaster. Indonesia is one of the tropical climate countries, with majority hydrometeorological disasters that continue to occur every end to the beginning of the year in the rainy season. In Indonesia, tornadoes were the most frequent disaster and caused damage to homes, while the victims of death and disappearance were mostly caused by floods [1]. Because of the magnitude of the impact caused by extreme rainfall, spatio-temporal modeling is needed to utilize the similarity of spatial characteristics to be able to predict extreme weather for unobserved locations and obtain temporal patterns from extreme climates.

DOI: 10.24507/icicel.14.01.89

Big data analysis such as climate modeling is a representation of complex phenomena, which may involve spatial and temporal interactions, and regional topography. The Bayes method is one solution in representing these complex phenomena by designing a hierarchical structure for data and its parameters. Some studies that use Bayes modeling in predicting weather include [2] modeled land fire data in Portugal with explanatory variables being wind direction and speed, vegetation and local topological conditions, while [3] used a dynamic linear model on monthly maximum wind speed data.

Ordinarily, Bayesian inference obtains and predicts posterior distribution by using the MCMC algorithm [4]; however, this algorithm has the convergence issue problem and inefficient for spatio-temporal models that are arranged hierarchically. INLA (integrated nested Laplace approximation) is a solution to the limitations of MCMC, which has recently been used and is still being developed. INLA is designed to improve the efficiency and accuracy of posterior distribution estimation by utilizing Laplace's approximation. [5] used the hierarchical Bayes method to model daily precipitation data in Norway with INLA inference, which is a new method developed to overcome convergence problems from MCMC inference [5-7].

The main objective of this paper is to predict quantile of monthly rainfall for observed and unobserved locations, using the generalized Bayes spatio-temporal model as used in [5]. Some improvisations are also carried out innovatively to get more accurate predictions such as: 1) spatial and temporal smoothing parameters are essential for borrowing strength across locations and efficiently estimating spatial and temporal trends; therefore, we revise the estimated value of spatial smoothing parameters with local regression method, and temporal smoothing parameter with random walk of order 2 method; 2) threshold, u, has an important role in assessing extreme data, a careful bias-variance assessment must be performed to fix a suitable threshold; therefore we revise u using measure of surprise (MoS) method as in [8,9].

From these improvisations, we obtained satisfying results including a more efficient model compared to the cross validation study in [5] and the good correlation and RMSEP mean value for spatio-temporal model. In the remainder of this paper, we present the dataset, the detailed methodology and INLA inference which are explained in Section 2. Results and discussions are reported in Section 3. Some concluding remarks and possible future development are summarized in Section 4.

2. **Dataset and Methodology.**

In this section, we present dataset, generalized Bayes spatio-temporal modeling with its improvisations, and Bayesian inference using INLA.

2.1. **Dataset.**

The complete dataset consisted of monthly rainfall accumulations recorded in milliliters at 57 stations during the period 1981-2017. The data were divided into training set which was made available to spatio-temporal model, and a validation set which was used to assess quantile predictions. In the training period, we have a mixed dataset comprised of 45 observed (rich) stations and 12 unobserved (poor) stations. At the poor stations, there are absolutely no observation or the number of samples $n = 0$. In validation set, the data varies greatly. Only 1 station that has full sample size $n = 144$, while as many 8 stations 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 province, Indonesia.

2.2. **Bayesian spatio-temporal model.**

We decompose space-time modeling into three stages:

Stage 1: Let Y_0^+ state the intensity of rainfall that is positive, i.e., $Y_0^+ = Y(s,t)|Y(s,t) > 0$ assumed to have a gamma distribution

$$Y_0^+ \sim Gamma\{y; \mu(s,t), k\} := \frac{k^k}{\mu(s,t)^k \Gamma(k)} y^{k-1} \exp\left\{-\frac{ky}{\mu(s,t)}\right\}, \quad y > 0. \tag{1}$$

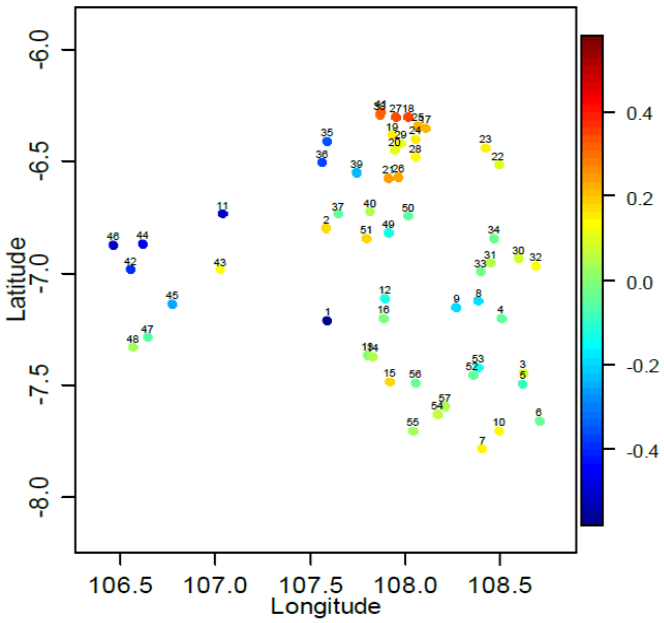

FIGURE 1. (color online) Map of monitoring locations, colored according to the estimated random effect

Stage 2: Threshold $u(s,t)$ is derived from MoS method, then we defined exceedance indicator as Bernouli's random variable which states that daily rainfall exceeds the threshold, i.e., $Z_u(s,t) = \mathbf{I}\{Y(s,t) > u(s,t)\}$,

$$Z_u(s,t) \sim Ber\left\{z; p_u(s,t)\right\} := p_u(s,t)^z \left\{1 - p_u(s,t)\right\}^{1-z}, \quad z \in \{0,1\}. \tag{2}$$

Stage 3: With $u(s,t)$ derived from stage 2, positive *exceedance* $Y_0^+ = Y(s,t) - u(s,t)|Y(s,t) - u(s,t)$ assumed to have reparameterized GP distribution, which is a function of q-quantile, $\kappa_q(s,t)$ and shape $\xi \geq 0$. Therefore, generaly, α-quantile, $y_\alpha(s,t)$ is

$$y_\alpha(s,t) = \begin{cases} u(s,t) + \kappa_q(s,t) \left[\left\{ \dfrac{1-\alpha}{p_u(s,t)} \right\}^{-\xi} - 1 \right] \Big/ \left\{ (1-q)^{-\xi} - 1 \right\}, & \xi \neq 0, \\ u(s,t) + \kappa_q(s,t) \log \left\{ \dfrac{1-\alpha}{p_u(s,t)} \right\} \Big/ \log(1-q), & \xi = 0. \end{cases} \tag{3}$$

To represent location and time diversity in spatio-temporal parameters in each step, a regression equation is formulated additively, which is the sum of the spatial and temporal random components which are assumed to be separable, as follows:

$$\log\left\{\mu(s,t)\right\} = \beta_0^{Gam} + x^{Gam}(s) + x^{Gam}(t), \tag{4}$$

$$\text{logit}\left\{p_u(s,t)\right\} = \beta_0^{Ber} + x^{Ber}(s) + x^{Ber}(t), \tag{5}$$

$$\log\left\{\kappa_q(s,t)\right\} = \log\left\{\mu(s,t)\right\} + \beta_0^{GP} + x^{GP}(s) + x^{GP}(t). \tag{6}$$

2.3. **Improvisations of spatial, temporal and threshold parameters.** In point data, spatial influence $x^{Gam}(s)$, $x^{Ber}(s)$ and $x^{GP}(s)$ are defined by the Matérn correlation function in [10],

$$Cov\left\{x(s_1), x(s_2)\right\} = \tau_s^{-1} \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}h}{\psi} \right)^v K_v \left(\frac{\sqrt{2v}h}{\psi} \right), \tag{7}$$

with $h = ||s_1 - s_2||$ being Euclidean distance, K_v with $v = 1$ is modified Bessel function and ψ is spatial range (smoothing) parameter which has an important role in borrowing

strength of spatial effects across nearby locations to predict the unobserved stations. We perform the range of ψ using local regression method according to [11], and we derived ψ about 106 km.

The temporal effects $x^{Gam}(t)$, $x^{Ber}(t)$ and $x^{GP}(t)$ are defined assumed to have a normal distribution, then $x^*(\omega+1) - 2x^*(\omega) + x^*(\omega-1) \sim Normal\left(0, \tau_t^{-1}\right)$. τ_t effect the temporal trends across locations; therefore, we obtained by modeling rainfall separately with temporal effects only as follows: $\log\{\mu(s,t)\} = x^{Gam}(t)$, so that we have τ_t being 0.035. We assume the temporal effects in monthly basis and annually cyclic.

Threshold, u, is a very important parameter in extreme data modeling using GP distribution. Determination of u values is a scheme to balance bias and variance of estimators. Too low u may cause bias in the estimators, while too high u implies a large estimation variance due to the small numbers of data that exceed the threshold [12-14]; therefore u selection must be performed carefully. In our application, we select u by MoS [8,9]. MoS is useful for calculating the degree of discrepancy between the data with the given distribution. The degree of incompatibility is measured by the expected surprise value close to 0.5, whereas a value close to 0 or 1 indicates u selection mismatch. The u for GP distribution is chosen when the surprise value converges to 0.5. For example, in Figure 2 the estimated u for station 20 is 171 millimeters, because 171 is the minimum point when the surprise value convergence around 0.5.

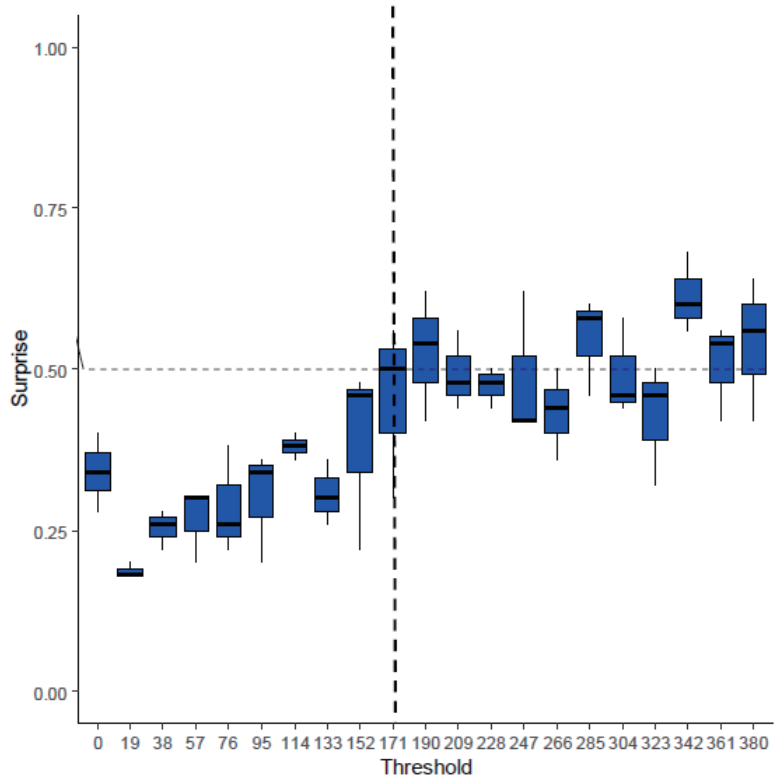

FIGURE 2. Threshold selection using MoS for station 20

2.4. **Bayesian inference with INLA.** Let $y(s_i,t_i) = (y_1, y_2, \ldots, y_m) = \mathbf{y}$, $i = 1, 2, \ldots, m$ is the observation data with the latent Gauss explanatory variable declared as $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)^T$ then $\eta_i = \beta_0 + x(s_i) + x(t_i)$, $\boldsymbol{\theta}_y$ is a vector for hyperparameters for y, and vector for hyperparameters for spatial and temporal random component is $\boldsymbol{\theta}_x$. The distribution of prior hyperparameters is defined as $\pi(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_x)$, and Gaussian probability $\mathbf{x}$ can be written as $\pi(\mathbf{x}|\boldsymbol{\theta}_x)$. Let $\pi(y_i|\eta_i, \boldsymbol{\theta}_y)$ be a *likelihood* from y_i with condition of the explanatory variables η_i and *likelihood* from hyperparameters $\boldsymbol{\theta}_y$.

INLA is an analytical Bayes-based inference, which can be applied to the generalized additive model that is complex and hierarchical and produces an approximation to the two posterior distributions of the following single variables:

$$\pi(\theta_k|y) = \int \pi\left(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}\right) dx d\boldsymbol{\theta}_{-k}, \tag{8}$$

$$\pi(x_i|\mathbf{y}) = \int\int \pi\left(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}\right) d\mathbf{x}_{-i} d\boldsymbol{\theta} = \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{9}$$

The Laplace approximation is applied nestedly, to determine the posterior distribution of the hyperparameter $\pi(\boldsymbol{\theta}|\mathbf{y})$ at (8), and the posterior distribution of parameters $\pi(x_i|\mathbf{y})$ at (9). More details on INLA estimation procedure and its statistical properties can be seen in [6,11,15].

3. **Main Results.** To predict monthly rainfall quantile in each station using generalized Bayes spatio-temporal model in Equations (4)-(6), we derived spatial and temporal random component and 95% pointwise credibility interval as in Figure 3. From Figure 3 (top), the model successfully predicts the spatial characteristics for observed and unobserved locations. The spatial random components have almost significance for observed locations. For poor station, the credibility interval is quite large, except for station 13, in which the random component is almost significance and the credibility interval is small rather than the other poor stations. It is because, station 13 is very close with the rich

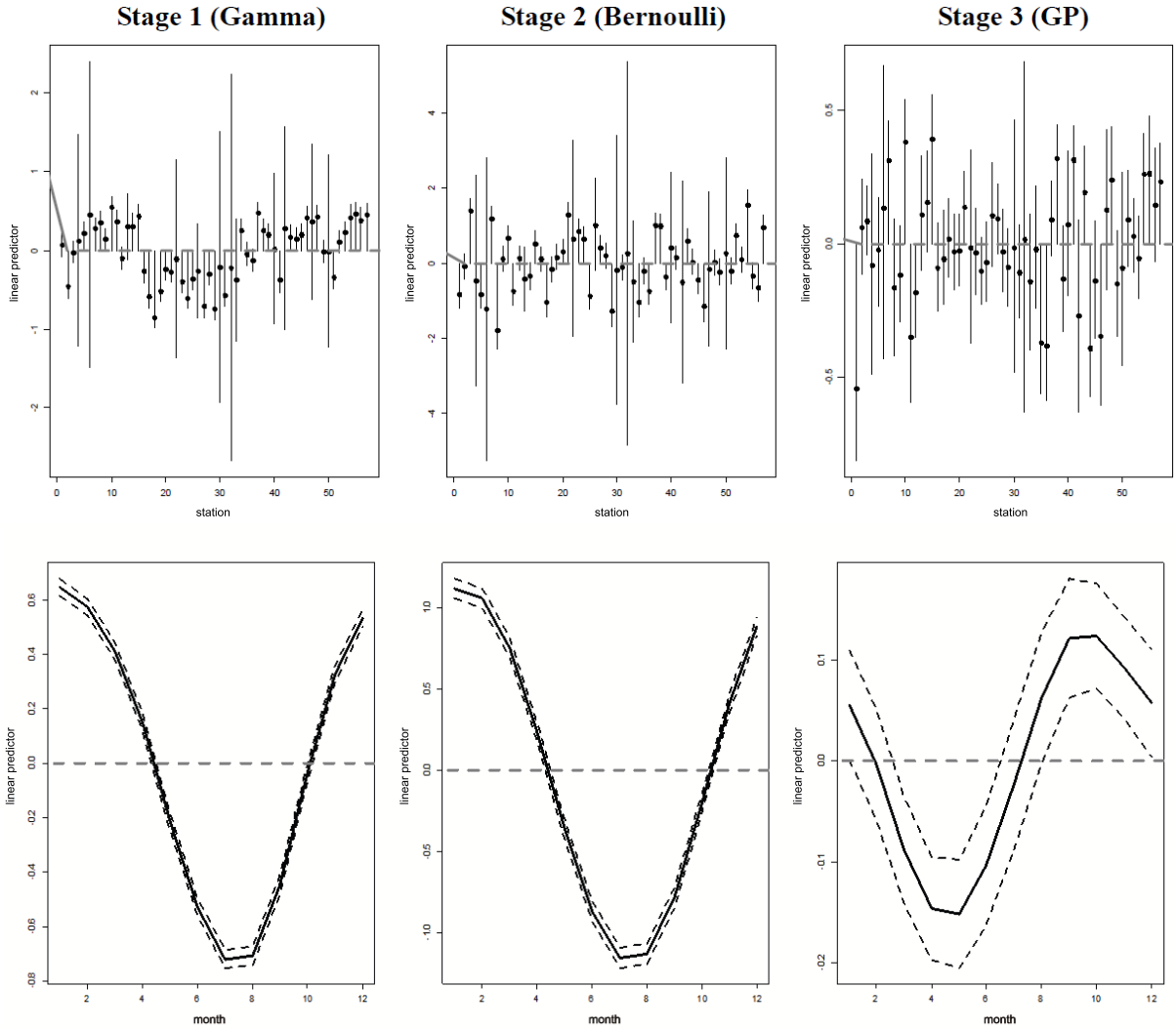

FIGURE 3. Spatial (top) and monthly (bottom) effects for the three stages

station 14. It means the purposed model successfully borrows the strength of rich stations to nearby locations. The estimated extreme spatial random effects for stage 3 (GP) are presented in Figure 1. The poor stations have unique spatial effects. The model shows that nearby locations have the same spatial characteristics.

We capture the temporal trend of monthly rainfall in Figure 3 (bottom). In stage 1 (Gamma), the monthly rainfall 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. In stage 2 (Bernoulli), identification of extreme rainfall has the same behavior as in stage 1, while in stage 3 (GP), the probability of extreme rainfall with significant positive effects increasingly occurs from May to August and decreases until the end of October. The estimated GP shape parameters for stage 3 has the positive posterior mean $\xi = 0.002$, with 95% credibility interval (0.0001, 0.006), showing that although the posterior mean is quite small, the effect is significance and the rainfall data are heavy tailed. We choose the prior of ξ using the concept of penalized complexity prior, for more details in [5,16].

Our goal is to predict quantile monthly rainfall for observed and unobserved locations from derived model parameters. Using validation data, the results are presented in Table 1. In validation period monthly rainfall is not perfectly recorded; this resulted in not all locations can be validated. Therefore, we determined the goodness of our model using the mean of correlation and RMSEP (root mean square error prediction) between predicted and real rainfall data and by setting aside unvalidated locations. At low (quantile 0.65) and moderate extreme (quantile 0.80) the correlation mean is around 0.8, while in high extreme (quantile 0.95 and 0.975) getting lower around 0.65-0.7. At the end of this study in Figure 4, we present the rainfall classification for all spatial locations in West Java using the classical non-parametric theory of local regression. According to resulted shape parameter $\xi = 0.002$ from stage 3 (GP), it shows that rainfall data have the upper extreme type, but this data does not show a high extreme corresponding to a small shape parameter value. Therefore, using 0.65 quantile, from Figure 4, in rainy season the west, east and southern parts of West Java tend to have higher rainfall than in the north. In dry season, it has an average rainfall that is almost evenly distributed throughout the West Java region.

Table 1. Estimated quantile of monthly rainfall

Quantile	Correlation mean	RMSEP mean
0.65	0.799	195.032
0.80	0.792	137.191
0.95	0.722	224.843
0.975	0.649	229.607

4. **Conclusions.** This paper combines three-stages modeling with three distributions, i.e., gamma, Bernoulli and generalized pareto (GP) distribution in extreme value theory with a flexible Bayesian approach to predict the amount of monthly rainfall for observed and unobserved locations. The purposed model successfully predicts even for the unobserved locations with good correlation and RMSEP mean overall. For spatio-temporal cases with many unobserved locations and imperfect validation data, our predicted quantile is good and could be made more complex if required by the context. We can enhance fixed effects like altitude or general circulation model (GCM) simulation data which is widely used in medium and long term weather prediction in statistical downscaling modeling as some of the following research [17-19]. Our improvisations for spatial, temporal and threshold selection as stated in Section 2.3 has succeeded in producing a model that is

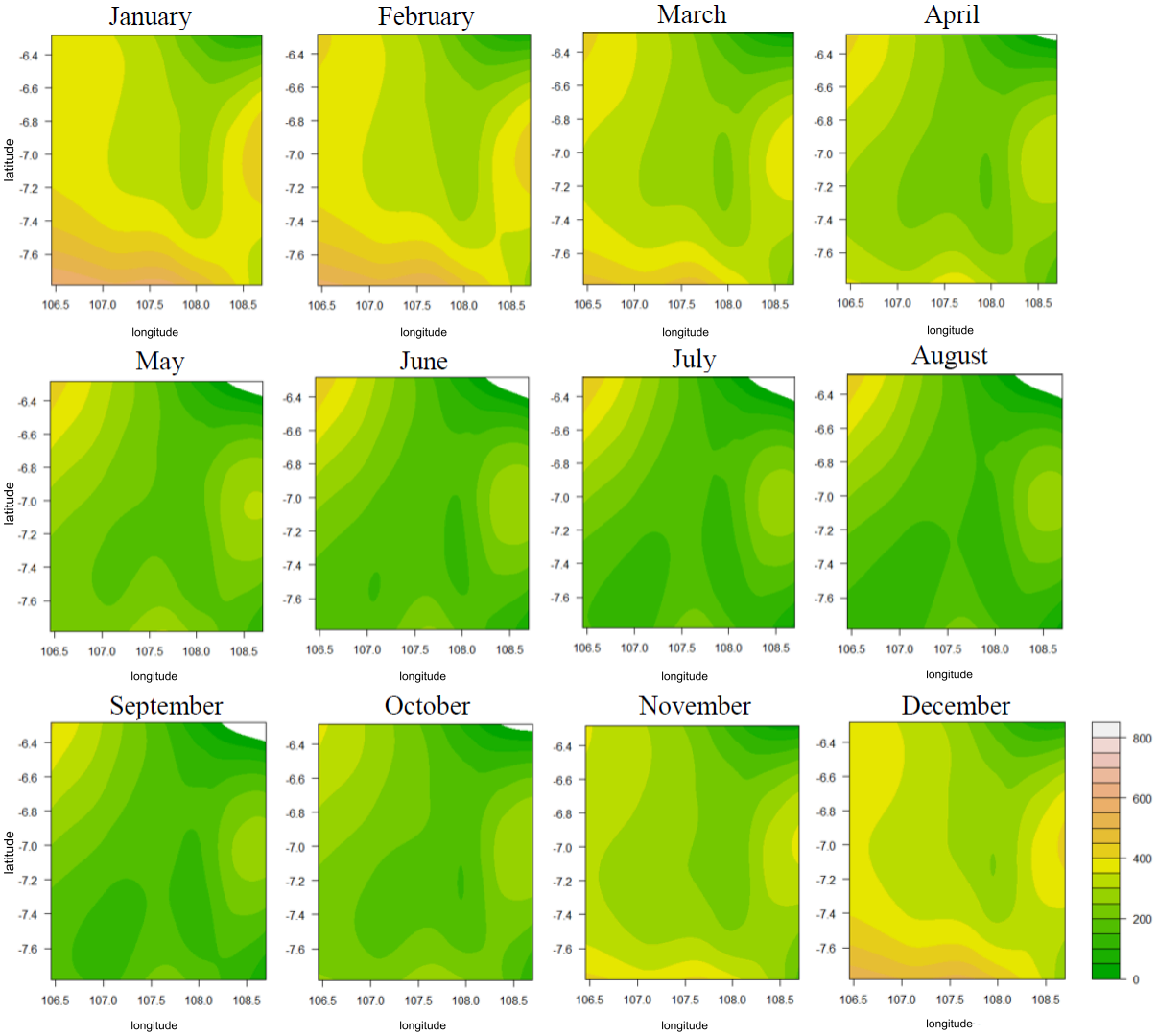

FIGURE 4. (color online) Regional monthly rainfall of West Java, Indonesia (0.65 quantile)

quite precise and more efficient, compared to cross validation studies in [5] which require high computational resources.

Acknowledgment. This work is fully supported by national grant from Kemenristek DIKTI (Kementrian Riset Teknologi dan Pendidikan Tinggi) of Indonesia.

REFERENCES

[1] *Indonesian National Board for Disaster Management (or Simply BNPB)*, https://bnpb.go.id/, 2019.
[2] K. F. Turkman, M. A. Turkman and J. Pereira, Asymptotic models and inference for extremes of spatio-temporal data, *Extremes*, vol.13, no.4, pp.375-397, 2010.
[3] B. Mahmoudian and M. Mohammadzadeh, A spatio-temporal dynamic regression model for extreme wind speeds, *Extremes*, vol.17, no.2, pp.221-245, 2014.
[4] C. Yang, J. Xu and Y. Li, Bayesian geoadditive modelling of climate extremes with nonparametric spatially varying temporal effects, *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol.36, no.12, pp.3975-3987, 2016.
[5] T. Opitz, R. Huser, H. Bakka and H. Rue, INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles, *Extremes*, vol.21, pp.441-462, 2018.
[6] H. Rue, S. Martino and N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B*, vol.71, no.2, pp.319-392, 2009.

[7] M. Blangiardo and M. Cameletti, *Spatial and Spatio-Temporal Bayesian Models with R-INLA*, John Wiley & Sons, 2015.
[8] J. Lee, Y. Fan and S. A. Sisson, Bayesian threshold selection for extremal models using measures of surprise, *Computational Statistics and Data Analysis*, vol.85, no.1, pp.84-99, 2014.
[9] A. Manurung, A. H. Wigena and A. Djuraidah, GPD threshold estimation using measure of surprise, *International Journal of Sciences: Basic and Applied Research*, vol.45, no.3, pp.16-25, 2018.
[10] F. Lindgren and H. Rue, Bayesian spatial modelling with R-INLA, *Journal of Statistical Software*, vol.63, no.19, 2015.
[11] F. Lindgren, H. Rue and J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B*, vol.73, no.4, pp.423-498, 2011.
[12] A. C. Davison, S. A. Padoan and M. Ribatet, Statistical modeling of spatial extremes, *Statistical Science*, vol.27, no.2, pp.161-186, 2012.
[13] P. J. Northrop and P. Jonathan, Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights (with discussion), *Environmetrics*, vol.22, no.7, pp.799-809, 2011.
[14] C. Scarrott and A. MacDonald, A review of extreme value threshold estimation and uncertainty quantification, *REVSTAT – Statistical Journal*, vol.10, no.1, pp.33-60, 2012.
[15] H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson and F. K. Lindgren, Bayesian computing with INLA: A review, *Annual Review of Statistics and Its Application*, vol.4, pp.395-421, 2017.
[16] D. Simpson, H. Rue, A. Riebler, T. G. Martins and S. H. Sørbye, Penalising model component complexity: A principled, practical approach to constructing priors, *Statistical Science*, vol.32, no.1, pp.1-28, 2017.
[17] J. Tang, X. Niu, S. Wang, H. Gao, X. Wang and J. Wu, Statistical downscaling and dynamical downscaling of regional climate in China: Present climate evaluations and future climate projections, *Journal of Geophysical Research: Atmospheres*, vol.121, pp.2110-2129, 2016.
[18] T. B. N. Cahyani, A. H. Wigena and A. Djuraidah, Quantile regression with elastic-net in statistical downscaling to predict extreme rainfall, *Global Journal of Pure and Applied Mathematics*, vol.12, no.4, pp.3517-3524, 2016.
[19] W. J. Sari, A. H. Wigena and A. Djuraidah, Quantile regression with functional principal component in statistical downscaling to predict extreme rainfall, *International Journal of Ecological Economics and Statistics*, vol.38, no.1, 2017.