International Journal of Innovative
Computing, Information and Control
Volume **17**, Number **1**, February **2021**

ICIC International ©2021 ISSN 1349-4198
pp. **259–273**

EXTREME DATA ANALYSIS USING SPATIO-TEMPORAL BAYES REGRESSION WITH INLA IN STATISTICAL DOWNSCALING MODEL

Anik Djuraidah[1], Ro'fah Nur Rachmawati[1,2], 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
{ anikdjuraidah; aji_hw; wayanma }@apps.ipb.ac.id

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

Received July 2020; revised December 2020

Abstract. *Statistical downscaling (SD) is modeling technique using global scale data in the grid form to predict local scale data such as rainfall. To this present, SD with Bayes frameworks applied to spatio-temporal cases still uses the Markov chain Monte Carlo (MCMC) algorithm which demands expensive computational capabilities. Therefore, we present spatio-temporal Bayes regression in SD model with efficient inference method, namely INLA (integrated nested Laplace approximation), to predict local extreme rainfall at unobserved locations. The modeling algorithm uses a combination of three distributions, i.e., gamma, Bernoulli, and generalized Pareto respectively for modeling average, identification of extreme, and prediction of extreme. For high accuracy of extreme prediction, we innovatively propose improvisations in determining important parameters, i.e., spatial smoothing, extreme value threshold, and tail index parameters. Big data analysis consists of the spatio-temporal monthly local rainfall from 57 locations, and the monthly precipitation general circulation model (GCM) explanatory variables with 5 × 8 grid dimensions observed from 1981-2017. The model successfully predicts the unobserved locations with strong correlations between predictive and validation values about 0.81-0.84 for low to moderate extreme rainfall, and 0.70-0.72 for high extreme rainfall. Based on the RMSEP value, the proposed model is the best method for estimating rainfall to high extreme levels compared to other spatio-temporal Bayes models with INLA inference.*
Keywords: Extreme rainfall prediction, Big data analysis, Principal component analysis, INLA (integrated nested Laplace approximation), GCM (global circulation model)

1. **Introduction.** Statistical downscaling (SD) is one of the downscaling techniques often used in climate modeling by utilizing global scale data to obtain conclusions at a local scale. Global scale data usually used as explanatory is GCM (global circulation model) output that represents a variety of systems found on earth, including the atmosphere, oceans, land surface and sea ice which are very useful for research on climate change and variability [1]. Big data analysis such as climate modeling is a representation of complex phenomena, which involve spatial, temporal (or its interaction), regional topography and

DOI: 10.24507/ijicic.17.01.259

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. SD modeling with Bayes frameworks applied to spatio-temporal data has a very large amount of observation, because the model collects data from various spatial locations and relatively long observation time intervals.

Extreme undesirable events such as hydrometeorological disasters, can have an enormous impact on material and life losses. Various prevention activities are strived to reduce the impact of losses incurred, and one of these is the prediction of extreme events. Estimation of extreme originated from the first theory of extreme value, that is block maxima which focuses on asymptotic behavior from maximum samples, which converges to the generalized extreme value (GEV) distribution. The block maxima method is very popularly used in the field of environmental science, for example, the distribution of GEV is used for annual maximum temperature data or maximum annual river discharge, see [2-5] for the current studies. The second theory of extreme values is peaks over threshold (POT), focusing on samples that are above the threshold (u), so that data X that exceeds u is quite high, $Y = X - u > 0$ converges to the generalized pareto (GP) distribution, see [6-9] for recent applications.

Ordinarily in the current research, the posterior distribution estimations of spatio-temporal Bayes inference in SD model still use MCMC. [2] uses MCMC to analyze annual minimum temperatures for the past 6 decades in China, while [3] uses a dynamic linear model on monthly maximum wind speed data. For complex spatio-temporal Bayes in SD model whose processes and parameters are designed with a hierarchical structure, the computational time needed by MCMC is very long. This inefficiency not only has effects on the time and computational resources but also on the problem of convergence of the posterior distribution produced. INLA (integrated nested Laplace approximation) is a solution to the limitations of MCMC, which has recently been used and is still being highly used and developed. INLA is designed to improve the efficiency and accuracy of posterior distribution estimation by utilizing Laplace's approximation [10-12]. The use of INLA inference in very complex models, such as Bayes spatio-temporal data with global scale GCM in SD modeling, ensures that the resulting estimators are convergent, accurate and efficient. Current study [6] uses the hierarchical Bayes method with INLA to model daily precipitation data in Norway. However, [6] has not used global scale data, so this research is not included in the SD modeling category. Therefore, INLA inference for spatio-temporal Bayes in SD modeling has not been applied to the current research.

The main objectives of this research are to obtain temporal patterns and predict quantile of monthly rainfall for observed and unobserved (no data have been recorded) locations, using spatio-temporal Bayes model as used in our latest work on [13]. However, in this research, we enhance the prediction to SD modeling which is unique to this paper based on the results of latest international publications. Some improvisations are also carried out innovatively to get more accurate predictions, for example, 1) spatial smoothing parameters are extremely important in borrowing characteristics across spatial locations and efficiently estimating spatial pattern. Therefore, we revise the estimated value of spatial smoothing parameters using classical method local regression; 2) threshold u has an important role in assessing extreme data. An appropriate threshold u must be determined carefully to assess bias variance trade-off; therefore, we revise u using measure of surprise (MoS) method as in [14,15]; 3) tail index parameter ξ is very influential in modeling the extreme distribution. For this reason, the value ξ should no longer be of constant value but by establishing a prior distribution. In this study, the parameter ξ uses the penalized complexity (PC) prior which was introduced by [16].

In the remaining part of this paper, we present the dataset description and its preprocessing in Section 2. The purposed spatio-temporal Bayes in SD model, the improvisations of spatial, extreme value threshold and tail index parameters, INLA inference and modeling algorithm are explained in Section 3. Results, discussions and comparisons with our works on [13,17] are reported in Section 4. Some concluding remarks and possible future developments are summarized in Section 5.

2. **Dataset Description.** The local response dataset consists of monthly rainfall recorded in milliliters at 57 stations in West Java, Indonesia, during the period 1981-2017. There are so many stations with missing data cases, and there are even stations that do not have observations (unobserved) at all, as happened at new stations. Estimation of rainfall for unobserved locations is a challenging statistical problem, this is because the more missing data, the greater the predictive bias value. Based on this background and as one of the important contributions in this study, the proposed model includes a spatial effect, so that unobserved locations can be estimated by borrowing spatial characters from other observed locations that are close to each other. The data were divided into training set (January 1981 – December 2005) which was made available to spatio-temporal Bayes SD model, and a validation set (January 2006 – December 2017) which was used to evaluate quantile predictions. To see the ability of the proposed model in estimating unobserved locations, in training period, a varied dataset is composed of 12 unobserved (stations 4, 6, 11, 22, 26, 30, 32, 33, 40, 42, 47 and 50) and 45 observed (the rest) stations. The exact coordinates of stations are shown in Figure 1. West Java is a province with very diverse rainfall, this is because the morphology of the region tends to form slopes, with the northern part bordering the Java Sea and the south with the Indian Ocean. In the rainy season, the local government establishes this area as an area prone to flooding and landslides [18].

The GCM data used is precipitation, which was issued by the National Centers for Environmental Prediction (NCEP) in the form of Climate Forecast System Reanalysis

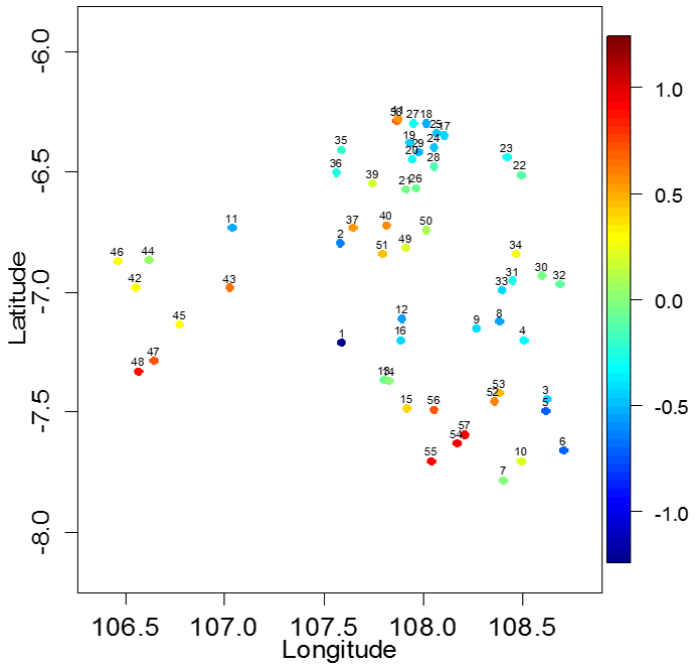

FIGURE 1. (color online) Map of monitoring locations, colored according to posterior distribution of spatial random component in Equation (8)

(CFSR). CFSR is a model that describes a global interaction between land, sea, and air on earth that is measured every 6 hours or 4 times a day. In this study the variables used were the average precipitation rate taken from the website https://rda.ucar.edu/. The average precipitation rate from the GCM data has a grid type and covers all of the West Java land, with grid size as $2.5^\circ \times 2.5^\circ$ and 5×8 dimensions (40 variables) as can be seen in Figure 2.

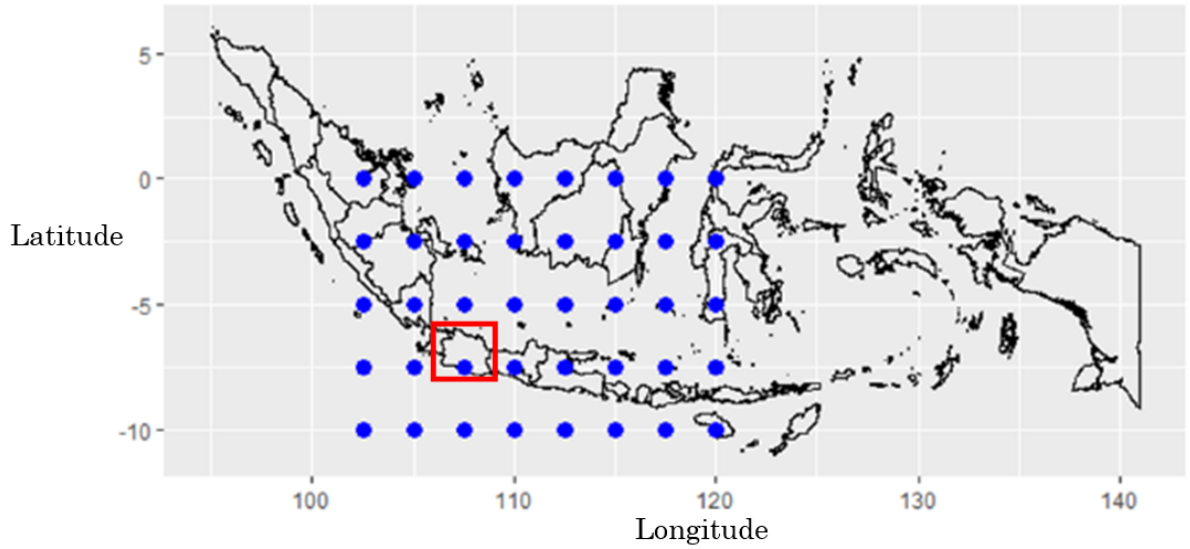

FIGURE 2. 5×8 grid GCM (blue dots), West Java (red rectangle), Indonesia

GCM provides the best tools for making climate change projections, but they are only designed to describe the large-scale features and do not provide sufficient details for many applications. However, the local climate is linked to the large scales, and it is possible to make some inferences about local climate change through downscaling [19]. Principal component analysis (PCA) is a statistical analysis that is still very often used in the handling of multicollinearity and the reduction dimensions of the GCM variable on statistical downscaling model. PCA reduces the dimensions of the data which have many variables that are correlated while maintaining variation of the initial data, so that the diversity of global variables can still be represented but also has a smaller dimension allowing for local scale predictions.

The PCA is used as a pre-processing technique to obtain latent variables that are orthogonal and are linear combinations of the covariates. The number of latent variables used for further analysis is generally determined by at least two of the following three things: scree-plot graphical representation, cumulative variance proportion, and the magnitude of the variance indicated by the Eigen value. According to [20] the graphical representation often leads to the first two or three principal components (PCs), cumulative variance proportion can be subjective with at least 70% is common and non-zero Eigen values. Therefore, this research uses 3 PCs, the cumulative variance proportion is taken $\geqslant 90\%$ and the Eigen value > 1. Cumulative variance proportion and Eigen value presented in Table 1 suggest to take as many as 3 principal components (PC) as orthogonal latent variables. The cumulative proportion for the first three PC of the GCM output has given a value of $\geqslant 90\%$. Variance values characterized by the Eigen values of the three main components have been obtained > 1. Therefore, further analysis uses three latent variables (PC1, PC2 and PC3 scores) for global explanatory variables in the purposed model.

TABLE 1. Eigen value and cumulative proportion

PC	Eigen value	Cumulative proportion
1	32.03	80.08
2	2.78	87.02
3	**1.17**	**90.00**
4	0.48	91.13
5	0.46	92.27

3. **Spatio-Temporal Bayes in SD Model.** We decompose rainfall modeling algorithm into space-time model combining three distributions which consist of 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

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

with $\mu(s,t)$ as mean parameter for location s at time t and k as shape parameter.

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

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

where $p_u(s,t)$ states the probability rainfall in location s at time t above the threshold u.

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. We know that $\mathrm{GP}(y;\xi,\sigma)$ is GP distribution with tail index (ξ) – scale (σ) parameters, so for $y>0$ the GP distribution is stated as follows:

$$F_{(\xi,\sigma)}(y) = \begin{cases} 1-\left(1+\dfrac{\xi y}{\sigma}\right)^{-1/\xi}, & \xi \neq 0 \\ 1-\exp\left(\dfrac{-y}{\sigma}\right), & \xi = 0 \end{cases}. \tag{3}$$

Using scale repameterization, i.e., $\sigma = \frac{-K_q}{\log(1-q)}$, the reparameterized $\mathrm{GP}(y;\xi,K_q)$ distribution can be written as

$$F_{(\xi,K_q)}(y) = \begin{cases} 1-\left(1+\left((1-q)^{-\xi}-1\right)\dfrac{y}{K_q}\right)^{-1/\xi}, & \xi \neq 0 \\ 1-(1-q)^{y/K_q}, & \xi = 0 \end{cases}, \tag{4}$$

which is a function of q-quantile, $\kappa_q(s,t)$ and tail index (or shape) $\xi \geqslant 0$. Therefore, generally, α-quantile, $\widehat{y}_\alpha(s,t)$ is

$$\widehat{y}_\alpha(s,t) = \begin{cases} u(s,t) + \kappa_q(s,t)\left[\left\{\dfrac{1-\alpha}{p_u(s,t)}\right\}^{-\xi} - 1\right] \Big/ \left\{(1-q)^{-\xi}-1\right\}, & \xi \neq 0 \\ u(s,t) + \kappa_q(s,t)\log\left\{\dfrac{1-\alpha}{p_u(s,t)}\right\} \Big/ \log(1-q), & \xi = 0 \end{cases}. \tag{5}$$

To represent location and time diversity in spatio-temporal parameters in each step, a regression equation is formulated additively, which is the sum of spatial and temporal

random components which are assumed to be separable. Therefore, the purposed spatio-temporal Bayes regressions in SD model are as follows:

$$\log\{\mu(s,t)\} = \beta_0^{Gam} + \beta_1 \mathbf{PC}_1 + \beta_2 \mathbf{PC}_2 + \beta_3 \mathbf{PC}_3, \tag{6}$$

$$\text{logit}\{p_u(s,t)\} = \beta_0^{Ber} + x^{Ber}(s) + x^{Ber}(t), \tag{7}$$

$$\log\{\kappa_q(s,t)\} = \log\{\mu(s,t)\} + \beta_0^{GP} + x^{GP}(s) + x^{GP}(t), \tag{8}$$

where β_0^{Gam}, β_0^{Ber}, and β_0^{GP} are the intercept regressions for each step, $x^{Ber}(s)$ and $x^{GP}(s)$ are spatial random components in steps 2 and 3, $x^{Ber}(t)$ and $x^{GP}(t)$ are temporal random components in steps 2 and 3, while β_1, β_2, and β_3 are the constant effects of principal components derived from dimension reduction of GCM data preprocessing.

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

$$Cov\left\{x(s_1), x(s_2)\right\} = \tau_s^{-1} \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}h}{\psi}\right)^v K_v \left(\frac{\sqrt{2v}h}{\psi}\right), \tag{9}$$

with $h = ||s_1 - s_2||$ as Euclidean distance, K_v with $v = 1$ as modified Bessel function and ψ as spatial range (smoothing) parameter which has an important role in borrowing strength of spatial effects across nearby locations to predict the unobserved stations. [10] states that the distance which represents spatial smoothing parameter can be determined so that the rainfall correlation between spatial locations is close to 0.1. We perform the range of ψ using local regression method according to [10], and we derive ψ is about 106 km.

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 [11,12,14]; therefore, u selection must be performed carefully. In our application, we select u by MoS [9]. MoS is useful for calculating the degree of discrepancy between the data with the given distribution. The degree of incompatibility measured by the expected surprise value is 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 3 the estimated u for station 20 is 171 millimeters, because 171 is the minimum point when the surprise value convergence around 0.5.

The distribution for the tail index ξ is an important part of modeling extreme values based on GP distribution. One distribution used to model the tail index is penalized complexity (PC) prior which was introduced by [16]. The PC prior gives a "penalty" for the reference model of the base model; in other words, the PC prior is designed to produce a simpler model, by giving penalties to more complex models. Penalties are given to reference models based on the concept of "distance" $d(f_\xi, f_{\xi_0})$, with f_ξ and f_{ξ_0} respectively being the reference and the base model.

The concept of distance used in [16] is based on Kullback-Leibler divergence (KLD), i.e.,

$$d\left(f_\xi, f_{\xi_0}\right) = \sqrt{2\text{KLD}\left(f_\xi || f_{\xi_0}\right)},$$

with

$$\text{KLD}\left(f_\xi || f_{\xi_0}\right) = \int f_\xi(y) \log \frac{f_\xi(y)}{f_{\xi_0}(y)} dy. \tag{10}$$

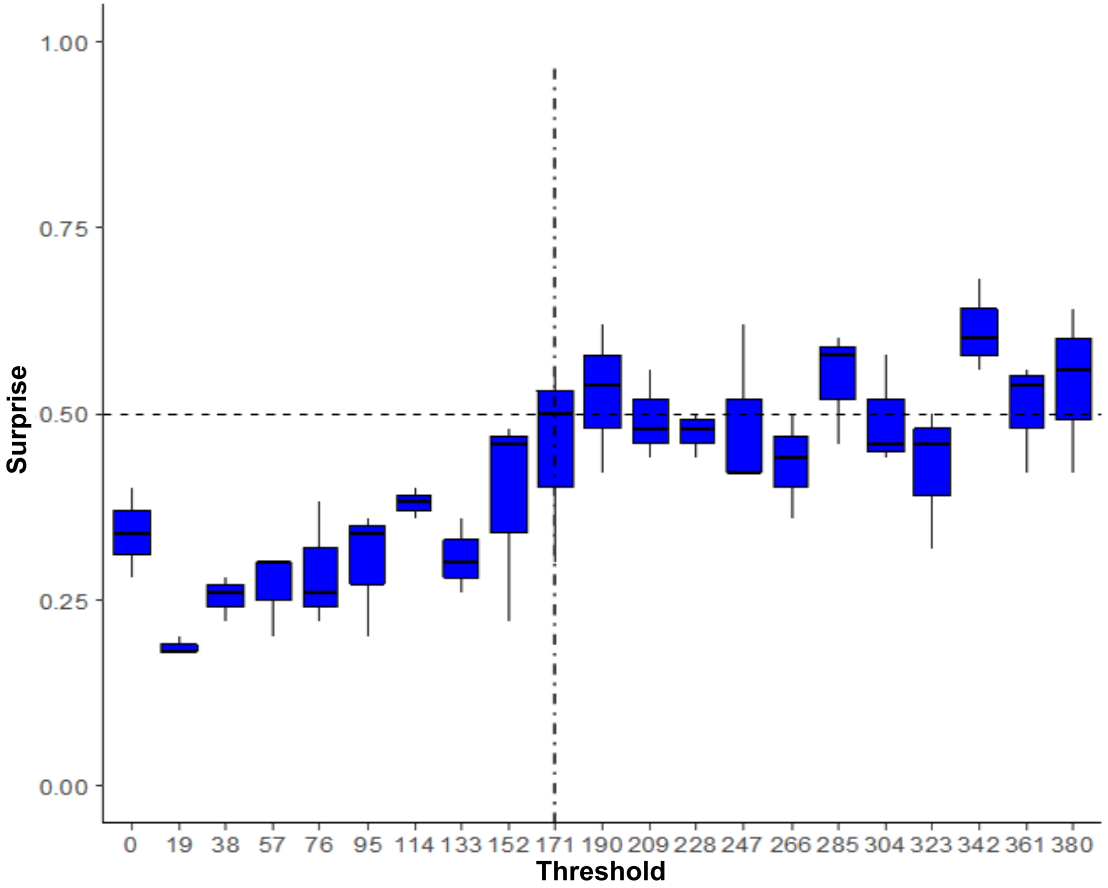

FIGURE 3. Threshold selection using MoS for station 20

For GP distribution as in (3), the reference model distribution for ξ is

$$f_\xi(y) := \frac{1}{\sigma}\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1}, \quad y>0,\ \sigma>0,\ \xi>0$$

and the distribution of the most natural base model in selecting prior distribution ξ is the exponential distribution, which is GP distribution for $\xi = 0$:

$$f_{\xi_0}(y) := \lim_{\xi\to 0} f_\xi(y) = \frac{1}{\sigma}\exp\left(\frac{-y}{\sigma}\right), \quad y>0,\ \sigma>0,\ \xi=0.$$

Therefore, KLD in Equation (10) can be obtained with

$$\mathrm{KLD}\left(f_\xi||f_{\xi_0}\right) = \int_0^\infty \frac{1}{\sigma}\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1} \log\left(\frac{\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1}}{\exp\left(\frac{-y}{\sigma}\right)}\right)dy.$$

By using $t = \frac{1}{\xi}\log\left(1+\frac{\xi y}{\sigma}\right)$, we obtain $dy = \sigma\left(1+\frac{\xi y}{\sigma}\right)dt$; thus, the above equation can be written as

$$\begin{aligned}\mathrm{KLD}\left(f_\xi||f_{\xi_0}\right) &= \int_0^\infty \frac{1}{\sigma}\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1}\left[\log\left\{\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1}\right\} - \frac{y}{\sigma}\right]dy \\ &= \int_0^\infty \frac{1}{\sigma}\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1}\left(-\frac{1}{\xi}-1\right)\log\left(1+\frac{\xi y}{\sigma}\right)dy - \int_0^\infty \frac{1}{\sigma}\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1}\frac{y}{\sigma}\end{aligned}$$

$$\begin{aligned}\mathrm{KLD}\left(f_\xi||f_{\xi_0}\right) &= \left(-\frac{1}{\xi}-1\right)\int_0^\infty \left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}} t\xi dt - \frac{1}{\sigma}\left(\frac{\sigma}{1-\xi}\right)\\ &= \frac{\xi^2}{1-\xi}, \quad 0 \leqslant \xi < 1. \end{aligned} \tag{11}$$

Thus, the penalty given for reference model is $d\left(f_\xi, f_{\xi_0}\right) = \sqrt{2\mathrm{KLD}} = \sqrt{2}\xi(1-\xi)^{-1/2}$. Based on model in Equation (11), for $\xi \to 1$ then $\mathrm{KLD}\left(f_\xi||f_{\xi_0}\right) \to \infty$, therefore, in forming a penalty for f_ξ can be obtained by determining the value of the basic model parameters that meet the exact solution in Equation (11) and the approximation of Equation (11) for the value $\xi \to 0$.

The PC prior distribution for ξ using the exact solution in Equation (11) is

$$\begin{aligned}\pi(\xi) &= \lambda \exp\left(-\lambda d\left(f_\xi, f_{\xi_0}\right)\right)\left|\frac{\partial d\left(f_\xi, f_{\xi_0}\right)}{\partial \xi}\right|\\ &= \tilde{\lambda}\exp\left(\tilde{\lambda}\frac{\xi}{1-\xi}\right)\left\{\frac{1-\xi/2}{(1-\xi)^{3/2}}\right\}, \quad 0 \leqslant \xi < 1, \end{aligned} \tag{12}$$

with $\tilde{\lambda} = \sqrt{2}\lambda$. The PC prior distribution for the approximation of Equation (11) for the value $\xi \to 0$ is $d\left(f_\xi, f_{\xi_0}\right) = \lim_{\xi\to 0}\sqrt{2\mathrm{KLD}} = \lim_{\xi\to 0}\sqrt{2\frac{\xi^2}{1-\xi}} = \sqrt{2}\xi$, $0 \leqslant \xi < \infty$. Thus the approximation of PC prior distribution for ξ is

$$\pi(\xi) = \lambda\left(-\lambda d\left(f_\xi, f_{\xi_0}\right)\right)\left|\frac{\partial d\left(f_\xi, f_{\xi_0}\right)}{\partial \xi}\right| = \tilde{\lambda}\exp\left(\tilde{\lambda}\xi\right), \quad 0 \leqslant \xi < \infty, \tag{13}$$

with $\tilde{\lambda} = \sqrt{2}\lambda$. Based on Equations (12) and (13), the PC prior distribution for ξ is an exponential distribution with rate $\tilde{\lambda} = \sqrt{2}\lambda$.

Furthermore, the determination of value $\lambda = \tilde{\lambda}/\sqrt{2}$ can be determined subjectively by the researcher. In this study the value of λ is chosen so that the curve of Equations (12) and (13) have the same shape. For various scenarios of λ values (the plot curve not shown here), the greater the value of λ, plot the two curves show the same result, so the value of $\lambda = 10.5$ is chosen.

3.2. **Bayesian inference with INLA and rainfall modeling algorithm.** Let $y(s_i, t_i) = (y_1, y_2, \ldots, y_m) = \mathbf{y}$, $i = 1, 2, \ldots, m$ be the observation data with the latent Gauss explanatory variable declared as $\eta = (\eta_1, \eta_2, \ldots, \eta_m)^T$, η_i in Equations (6)-(8), $\boldsymbol{\theta}_y$ 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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_x)$, with Gaussian probability $\mathbf{x}$ being written as $\pi(\mathbf{x}|\boldsymbol{\theta}_x)$. Let $\pi\left(y_i|\eta_i, \boldsymbol{\theta}_y\right)$ be a likelihood from y_i with condition of the explanatory variables η_i and *likelihood* from hyperparameters $\boldsymbol{\theta}_y$. In our case vector $\mathbf{x}$ consists of intercept β_0^{Gam}, β_0^{Ber}, β_0^{GP}, β_i for $i = 1, 2, 3$, $x^{Ber}(s)$, $x^{GP}(s)$, $x^{Ber}(t)$ and $x^{GP}(t)$. Hyperparameter vector $\boldsymbol{\theta}_x$ consists of Matérn precision τ_s and precision for RW of order $2\tau_t$. In this paper we assume β_0^{Gam}, β_0^{Ber}, β_0^{GP}, β_i, for $i = 1, 2, 3 \sim \mathrm{Normal}\left(0, 10^5\right)$, and shape parameter $k \sim \mathrm{Gamma}(2, 2)$.

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:

$$\pi(\boldsymbol{\theta}_k|\mathbf{y}) = \int \pi\left(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}\right) dx d\boldsymbol{\theta}_{-k}, \tag{14}$$

$$\pi(x_i|\mathbf{y}) = \int\int \pi\left(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}\right) d\mathbf{x}_{-i} d\boldsymbol{\theta} = \int \pi\left(x_i|\boldsymbol{\theta}, \mathbf{y}\right) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{15}$$

The Laplace approximation is applied nestedly, to determining the posterior distribution of the hyperparameter $\pi(\boldsymbol{\theta}|\mathbf{y})$ at (14), and the posterior distribution of parameters $\pi(x_i|\mathbf{y})$ at (15). For more details on INLA estimation procedure and its statistical properties can be seen in [10,14,15]. Algorithm 1 presents the pseudocodes of our algorithm in modeling extreme rainfall using spatio-temporal Bayes regression in SD model. The α-quantiles, are used to compare the quantile values $\widehat{y}_\alpha(s,t)$ of low intensity of rainfall ($\alpha = 0.65$), moderate rainfall ($\alpha = 0.80$) and high extreme rainfall ($\alpha = 0.95, 0.975$) predictions.

Algorithm 1. Rainfall modeling

Input: The training data $\mathbf{Y}_{\text{train}}$, the test data $\mathbf{Y}_{\text{test}}$, assumption of parameters and hyperparameters

Output: The RMSEP (root mean square error prediction) and correlation between $\widehat{y}_\alpha(s,t)$ and $y_\alpha(s,t)$

for $\mathbf{Y}_{\text{train}}$ **do**
  Compute β_0^{Gam}, β_i for $i = 1, 2, 3$ in Equation (6)
  Compute threshold u for each location, using MoS method
  Compute β_0^{Ber}, $x^{Ber}(s)$ and $x^{Ber}(t)$ in Equation (7)
  Compute β_0^{GP}, $x^{GP}(s)$ and $x^{GP}(t)$ in Equation (8)
end for
Plot spatial and temporal random effect $x^{Ber}(s)$, $x^{Ber}(t)$ in Equation (7)
Plot spatial and temporal random effect $x^{GP}(s)$, $x^{GP}(t)$ in Equation (8)
for $\alpha = 0.65, 0.80, 0.95, 0.975$ **do**
  Compute $\widehat{y}_\alpha(s,t)$ in Equation (5)
end for
for $\mathbf{Y}_{\text{test}}$ **do**
  for each location s and time t **do**
    for $\alpha = 0.65, 0.80, 0.95, 0.975$ **do**
      Compute $y_\alpha(s,t)$
    end for
  end for
end for
Compute RMSEP and correlation between $\widehat{y}_\alpha(s,t)$ and $y_\alpha(s,t)$
return: RMSEP and correlation

4. **Results and Discussions.** The results of estimating the intercept value β_0^{Gam} and the coefficient of explanatory variable (the first three PC's score) β_i for $i = 1, 2, 3$ in Equation (6) are presented in Table 2. Intercepts and all explanatory variables have a significant influence on the estimation of average monthly rainfall at each location. This can be seen from the value of the credibility interval which does not contain 0. From this result, we have succeeded in determining the exact number of PC. This is because if the PCs are chosen correctly, then these variables will have a significant effect. And the existence of a significant effect indicates that the global fixed component, represented by the three PCs selected, plays an important role in estimating average rainfall.

The spatial random components $x^{Ber}(s)$, $x^{GP}(s)$ for Bernoulli and GP distribution based on Equations (7) and (8) are presented in Figure 4. Black circles are the posterior mean for estimated spatial random components, while blue lines represent 95% credibility

TABLE 2. Estimator of fixed effects in Equation (6)

Fixed effect	Mean	Standard deviation	Credibility interval
β_0^{Gam}	5.60	0.02	$(5.56, 5.65)$
β_1	0.05	0.01	$(0.04, 0.06)$
β_2	0.02	0.01	$(0.01, 0.03)$
β_3	-0.10	0.01	$(-0.12, -0.08)$

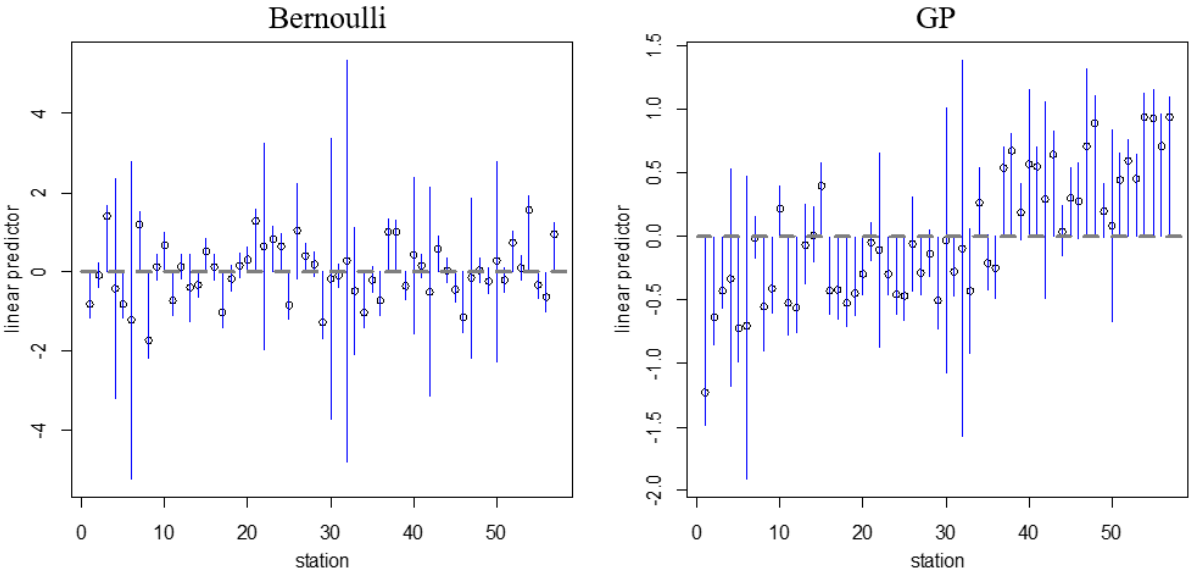

FIGURE 4. Posterior mean and credibility interval for spatial random components on the Bernoulli and GP distributions

intervals. From Figure 4 spatio-temporal Bayes regression in SD model can infer spatial characteristics especially for unobserved location. In general, the spatial random component for observed location has a significant value and its credibility intervals are shorter than those in unobserved locations. Location 32 is an unobserved station that has the greatest interval credibility for both Bernoulli and GP distribution. This is because the location is located in eastern part of West Java and is not near or surrounded by any observed station. Longer credibility intervals also occur in other unobservable locations, except for location 13 which have shorter credibility intervals. This is because the location is adjacent to the observed location 14.

The closeness extreme rainfall characteristics among locations can be represented by the mean posterior spatial random component, $x^{GP}(s)$ derived from the GP distribution and is presented in Figure 1. It can be seen that locations which are located close to each other have almost the same extreme rainfall characteristics. The lower extreme rainfall (below average 0) more scattered in the north and eastern part of West Java, while higher extreme rainfall (above average 0) is spread in west and southern part of West Java region.

The temporal random components $x^{Ber}(t)$, $x^{GP}(t)$ for the Bernoulli and GP distributions based on Equations (7) and (8) are presented in Figure 5. The single inner and two outer curve represents the posterior mean and 95% credibility interval respectively. The temporal random component with an annually cyclic behavior has a significant effect and produces an almost similar pattern in Bernoulli and GP distribution. During the rainy season, Bernoulli distribution identifies the extreme positive rainfall that began to occur in November and continues to increase until January. The highest identification of

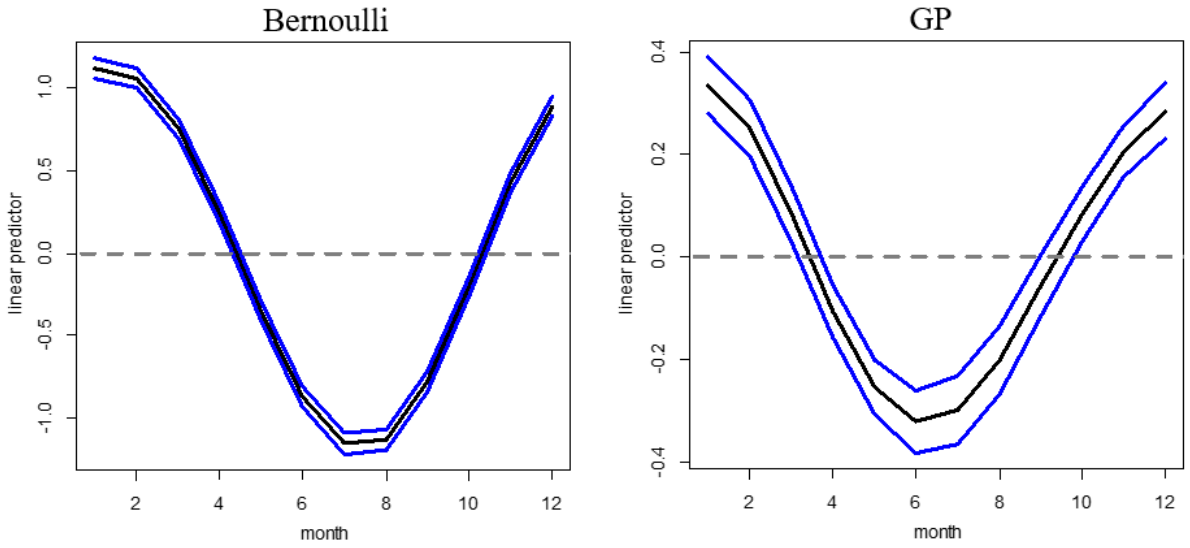

FIGURE 5. Posterior mean and credibility interval for temporal random components on the Bernoulli and GP distributions

extreme rainfall occurred in January and decreased from April to May. The dry season is indicated by the value of a negative temporal random component which indicates a decrease in identification of extreme rainfall, which occurs in May to September. On the GP distribution the intensity of extreme positive rainfall starts in October and continues to increase until January. January has the highest extreme rainfall, then decrease occurs until March, and enters the dry season from April to September. The estimated value of the tail index parameter on the GP distribution has a positive posterior mean $\xi = 0.002$ with credibility intervals $(0.0001, 0.0006)$. This indicates that although ξ value is quite small, it has a significant effect on estimating extreme rainfall, and a positive ξ indicates that the rainfall characteristics in 57 locations in the West Java region have the right tail distribution.

Model evaluations using RMSEP values and correlations for quantiles 0.65, 0.80, 0.95 and 0.975 were obtained using validation data and are presented in Table 3. It can be seen that the higher the quantile, the smaller the correlation and the greater the RMSEP. The best RMSEP and correlation are in the estimation of low extreme rainfall, that is at 0.65 quantile, with the smallest RMSEP about 116 mm and strong correlation of 0.84. Estimation of high extreme rainfall at 0.95 and 0.975 quantiles results in large RMSEP values, this is likely due to the very small amount of extreme data at quantiles 0.95 or more.

TABLE 3. RMSEP statistical value and correlation

Quantile	RMSEP	Correlation
0.65	116.55	0.84
0.80	132.51	0.81
0.95	217.32	0.72
0.975	281.30	0.70

To evaluate the performance of our purposed spatio-temporal Bayes regressions in SD model, we compare RMSEP and correlation criteria with our latest work on spatio-temporal Bayes regression in [13] and [17]. The details of the differences of the three spatio-temporal Bayes models are summarized in Table 4. Model in [13] becomes the

TABLE 4. Differences in spatio-temporal Bayes models

Model	Spatio-temporal Bayes specification	Assumption of Y	Include SD modeling?
Model in [13]	$\log\{\mu(s,t)\} = \beta_0^{Gam} + x^{Gam}(s) + x^{Gam}(t)$ $\text{logit}\{p_u(s,t)\} = \beta_0^{Ber} + x^{Ber}(s) + x^{Ber}(t)$ $\log\{\kappa_q(s,t)\} = \log\{\mu(s,t)\} + \beta_0^{GP} + x^{GP}(s) + x^{GP}(t)$	gamma, Bernoulli, GP	No
Model in [17]	$\log\{\mu(s,t)\} = \beta_0^{Nor} + x^{Nor}(s) + x^{Nor}(t) + \beta_1\text{PC}_1 + \beta_2\text{PC}_2 + \beta_3\text{PC}_3$	Normal	Yes
Purposed model	$\log\{\mu(s,t)\} = \beta_0^{Gam} + \beta_1\text{PC}_1 + \beta_2\text{PC}_2 + \beta_3\text{PC}_3$ $\text{logit}\{p_u(s,t)\} = \beta_0^{Ber} + x^{Ber}(s) + x^{Ber}(t)$ $\log\{\kappa_q(s,t)\} = \log\{\mu(s,t)\} + \beta_0^{GP} + x^{GP}(s) + x^{GP}(t)$	gamma, Bernoulli, GP	Yes

basis from our purposed model in this paper by developing average rainfall modeling in gamma distribution to SD model. Whereas [17] has included SD model but the rainfall is still using the assumption of normal distribution.

The comparison of RMSEP statistical values of spatio-temporal Bayes models in estimating extreme rainfall can be summarized in Figure 6 (top). The purposed model is generally the best method for estimating low (quantile 0.65) to high (quantile 0.95) extreme rainfall at 57 locations in West Java. The development of SD model on gamma distribution resulted in an RMSEP value that was much smaller than model in [13]. This shows that the score of three PC's selected from GCM output data has a significant influence in producing more accurate estimation in predicting low extreme rainfall. However, we found very different results on SD modeling based on normal distribution in model [17]. RMSEP generated for estimating moderate (quantile 0.80) to high extreme rainfall (quantile 0.95 and 0.975) is much higher than the other two models. In other words, spatio-temporal Bayes SD model with normal distribution is only able to predict low extreme rainfall. In general, the results in Figure 6 also show that our purpose model and the addition of GCM always give a smaller RMSEP compared to other approaches up to quantile 0.80.

Figure 6 (bottom) summarizes the comparison of correlations from spatio-temporal Bayes models in estimating extreme rainfall. It can be seen that spatio-temporal Bayes without involving SD modeling produces the lowest correlation value. While the proposed model produces the highest correlation up to the estimation of moderate extreme rainfall. While spatio-temporal Bayes model with normal distribution has the highest correlation in the estimation of high extreme rainfall. Therefore, based on the smallest RMSEP and the largest correlation criteria, generally, SD modeling can provide better estimation results compared to model without involving the influence of global scale data as explanatory variables.

5. **Conclusions.** This paper uses three combinations of distribution, i.e., gamma, Bernoulli and GP which were developed into SD modeling to obtain a temporal cycle and predict extreme rainfall at observed and unobserved locations. The purposed spatio-temporal Bayes in SD model is able to capture clearly and significantly annual rainfall patterns both in rainy and dry seasons. Our proposed model is also able to predict extreme rainfall even for unobserved locations with satisfactory RMSEP and correlation value. Improvisation of substantial parameters described in Section 3 has been successful

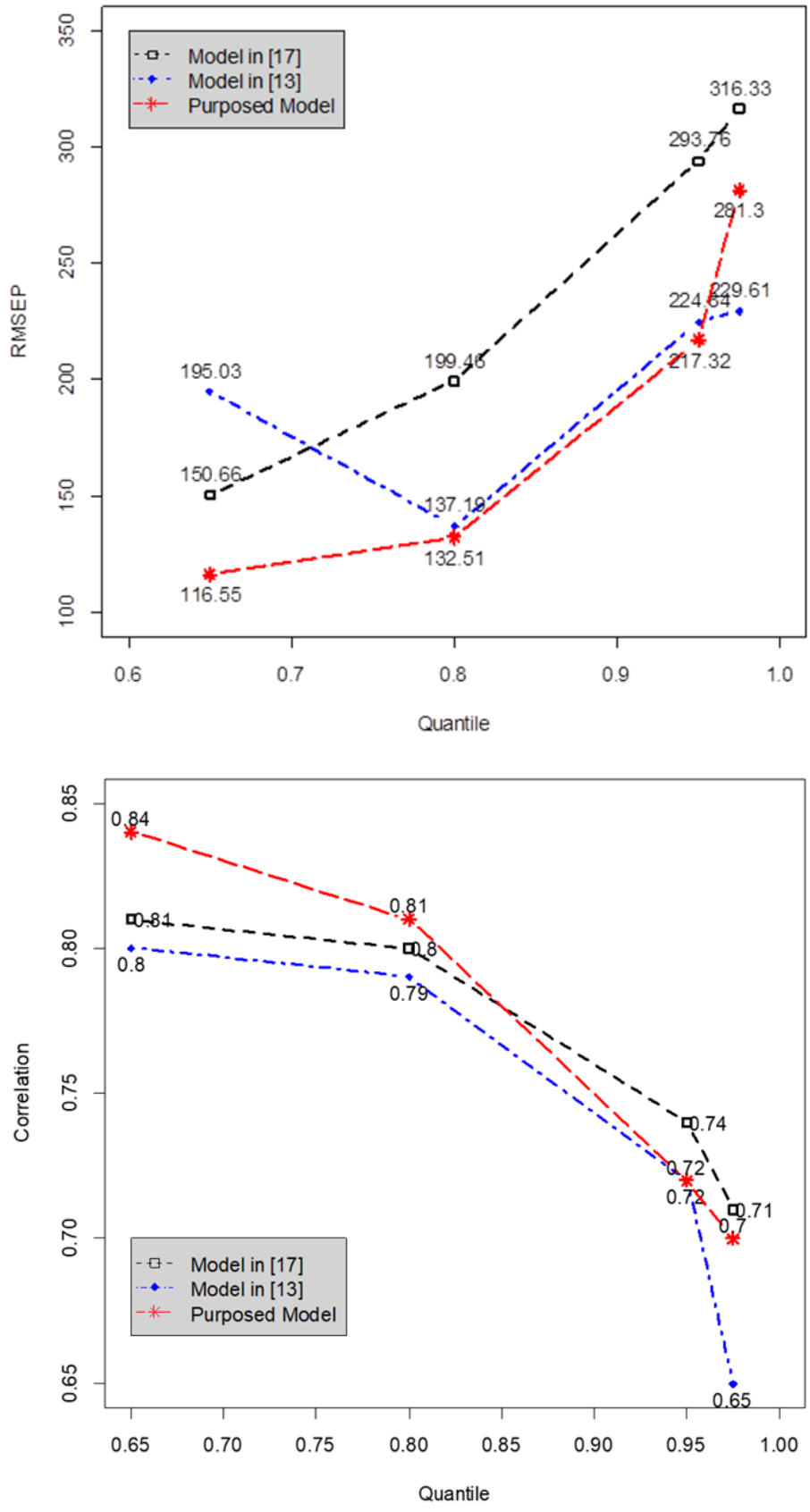

FIGURE 6. RMSEP and correlation of spatio-temporal Bayes models

in producing rainfall predictions that are more accurate and efficient than the cross-validation method in [6]. In this paper, development of SD model on gamma distribution is able to reduce RMSEP value significantly compared to the model in [13]. Generally, the proposed model also has a higher correlation value to moderate extreme rainfall prediction.

Further research can be carried out by developing the following aspects: 1) rainfall zoning. Estimation of extreme rainfall produced only applies to certain locations with specific coordinates. Then how about other locations that are not included in the observation. For this reason, the rainfall zoning is interesting to obtain characteristics of rainfall in the entire region. Zoning can be explored by the smoothing method as in [4], or by using the finite element method [21,22]; 2) rainfall projections. The advantage of SD modeling is essential for making future climate projections, so it can identify short, medium-to-long-term climate change, such as the following studies [23-25]. Rainfall projections that include the effects of climate variability and oceanographic events in a wider geographic area such as the El Niño and La Niña, are also very interesting research opportunities such as in [26]; 3) comparison with machine learning algorithm. Climate data that has been collected for decades from various places and types is a "Big Data". Not only in the form of text or raw data, but also radar and satellite imagery to video animation such as tropical cyclone movements, it requires high performance computing (HPC) to process it into a new climate information without adding it to the database. So even though the model we propose is very useful in predicting rainfall with missing or even unobserved data, comparing the prediction accuracy with machine (and/or deep) learning techniques is very interesting to study as in the following researches [27-29].

Acknowledgment. This work is fully supported by Lembaga Pengelola Dana Pendidikan (LPDP) by Ministry of Finance Republic of Indonesia and Grant of the Ministry of Research and Technology Republic of Indonesia in 2019. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] T. Yang, H. Li, W. Wang, C. Y. Xu and Z. Yu, Statistical downscaling of extreme daily precipitation, evaporation, and temperature and construction of future scenarios, *Hydrological Processes*, vol.26, pp.3510-3523, 2012.

[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] A. Bücher, J. Lilienthal, P. Kinsvater and R. Fried, Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis, *Extremes*, vol.23, no.2, 2020.

[6] 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.

[7] 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.

[8] M. Blangiardo and M. Cameletti, *Spatial and Spatio-Temporal Bayesian Models with R-INLA*, John Wiley & Sons, 2015.

[9] 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.

[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, 2016.
[13] R. N. Rachmawati, A. Djuraidah, A. H. Wigena and I W. Mangku, Extreme data analysis using generalized Bayes spatio-temporal model with INLA for extreme rainfall prediction, *ICIC Express Letters*, vol.14, no.1, pp.89-96, 2020.
[14] 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.
[15] 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.
[16] D. P. 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] R. N. Rachmawati, A. Djuraidah, A. H. Wigena and I W. Mangku, Spatio-temporal Bayes regression with INLA in statistical downscaling modeling for estimating West Java rainfall, *Proc. of the 1st International Conf. on Statistics and Analytics*, Bogor, Indonesia, pp.120-128, 2019.
[18] *Indonesian National Board for Disaster Management Website*, https://bnpb.go.id/, 2020.
[19] R. E. Benestad, D. Chen, A. Mezghani, L. Fan and K. Parding, On using principal components to represent stations in empirical statistical downscaling, *Tellus A: Dynamic Meteorology and Oceanography*, vol.67, no.1, 2015.
[20] I. T. Jolliffe and J. Cadima, Principal component analysis: A review and recent developments, *Philosophical Trans. Royal Society A*, vol.374, 2016.
[21] L. M. David, R. Glennie and A. E. Seaton, Understanding the stochastic partial differential equation approach to smoothing, *Journal of Agricultural, Biological and Environmental Statistics*, vol.25, pp.1-16, 2019.
[22] E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren and H. Rue, *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*, CRC Press/Taylor and Francis Group, 2019.
[23] E. M. Linder and E. Y. Pan, Statistical downscaling of regional climate model output to achieve projections of precipitation extremes, *Weather and Climate Extremes*, vol.12, pp.15-23, 2016.
[24] J. Patrick, Clemins, B. Gabriela, M. Jonathan, Winter, B. Beckage, E. Towler, A. Betts, R. Cummings and H. C. Queiroz, An analog approach for weather estimation using climate projections and reanalysis data, *Journal of Applied Meteorology and Climatology*, vol.58, no.8, pp.1763-1777, 2019.
[25] T. Mahtsente, Tadese, L. Kumar and R. Koech, Climate change projections in the Awash river basin of Ethiopia using global and regional climate models, *International Journal of Climatology*, vol.40, no.8, pp.3649-3666, 2020.
[26] K. Taweesin and U. Seeboonruang, The relationship between the climatic indices and the rainfall fluctuation in the lower central plain of Thailand, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.107-127, 2019.
[27] M. S. Balamurugan and R. Manojkumar, Study of short-term rain forecasting using machine learning based approach, *Wireless Network*, 2019.
[28] A. A. Shafin, Machine learning approach to forecast average weather temperature of Bangladesh, *Global Journal of Computer Science and Technology*, vol.19, no.3, 2019.
[29] A. Chattopadhyay, E. Nabizadeh and P. Hassanzadeh, Analog forecasting of extreme-causing weather patterns using deep learning, *Journal of Advances in Modeling Earth Systems*, vol.12, no.2, 2020.