Rainfall Prediction in East Java Using Spatial Extreme Value Theory

A A Artha¹ and A Sofro¹
¹Mathematics Department, Universitas Negeri Surabaya, Ketintang, Surabaya 60231, East Java, Indonesia

E-mail: ayuninsofro@unesa.ac.id

Abstract. Extreme rainfall is a phenomenon that occurs because of the high intensity of rainfall which causes many negative impacts. For example, is flooding that can ruin crops for farmers. To minimize this hazard, rainfall prediction is carried out. One methods that can be used is spatial extreme value theory using max stable approach and copula approach. From it, we will get the return level (predictive value) rainfall by paying attention to the location elements in it. This study discusses the max-stable approach using the Smith model and the copula approach using the Copula Gaussian model. Parameter estimation used is Maximum Likelihood Estimation (MLE) method. Because it takes into spatial elements, the trend surface model is also used. At the end of the study, Akaike Information Criterion (AIC) is used to compare and select the best model that can be used in predicting rainfall. Spatial extreme value theory will be applied to rainfall data in East Java which consists of nine rainfall observation stations. The results of spatial parameter estimation show that the highest of rainfall intensity value will occur around rainfall observation station namely Sawahan which is located in Nganjuk district in a period of two years. From AIC, it can be said that the model using copula Gaussian is the best model compare with max-stable approach as it provides the smallest AIC values.

1. Introduction
In this time, the condition of the earth became unstable. There are many disasters that causes many casualties, road damage, floods, and the other material losses. Natural disasters that occur come from unusual earth activities. As an example of the occurrence of floods, landslides, or rainstorms in an area caused by rainfall with high intensity. To anticipate this problem, predictions of extreme rainfall can be made. Natural events such as high rainfall, high tides, high winds, and etc occur because they are influenced by conditions from a place (spatial). Therefore, rainfall in one area will depend on rainfall in other areas which are close together due to spatial dependence [1]. R. Jane in 2016 used Copula Dependency on Radio Waves, Davison and Gholamrezae in 2010 used max-stable application for temperature data and Mc.Neil in 2013 used modelling dependencies with copula. So to predict extreme rainfall, spatial extreme value theory can be used with the max-stable approach and the copula approach because the approaches enable to resolve dependencies in many locations [1]. The max-stable approach and the copula approach are statistical methods used to analyze extreme events that occur on earth, such as rainfall, wind speed, humidity, and so on. Then forecasting and modeling extreme rainfall can be done by using the parameter estimation that is maximum likelihood estimation (MLE) [2]. This study takes the issue of rainfall in East Java. East Java is one of the provinces that calculated in the contributing of rice
production nationally. Around 17% of national rice production comes from East Java (BPS, 2016). Meanwhile, based on the East Java Agriculture Service in the first quarter of 2010 rice fields were damaged due to the impact of flooding reaching a significant value of 6,972.49 ha. So the rainfall in East Java is an important thing to be examined how rainfall patterns occur. The important thing in the studying of EVT is determine the return level which is the maximum rainfall threshold value [3]. So, the return level is an extreme rainfall prediction value that may occur in the future so that it can be used as a source of information that can be used to anticipate and minimize the impact.

2. Literature Review
2.1 Extreme Value Theory
Extreme Value Theory (EVT) is a theory that studies the occurrence of extreme values that occur in the earth. Identification of extreme values with EVT can be done with two methods that known as the Block Maxima (BM) method and the Peaks Over Threshold (POT) method. The BM method is a method which takes the maximum value in a period which called block and the POT method is a method where a value will pass through a certain threshold [4]. So to determine the extreme value from a data, BM method and POT method can be used [5]. In this study, BM method will be used because rainfall data is a seasonal data type where data patterns are influenced by season. BM method is a method that can identify extreme values based on the highest values of observation data. The block maxima method applies the Fisher and Tippett (1928) theorem in Gilli and Kellezi (2006), where in the theorem states that sample data of extreme values taken by the BM method will follow the Generalized Extreme Value (GEV) distribution function. Suppose there are $X_1, X_2, \ldots, X_n$ for each $X_n$ is a random variable that follows a distribution function $F_n$. So the $M_n$ distribution follows the cdf from GEV like the following equation:

$$P_r\left(\frac{M_n - B_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \rightarrow F(\mu, \sigma, \xi), \text{ as } n \rightarrow \infty,$$

where GEV has the cumulative distribution function (cdf) as follows:

$$F(\mu, \sigma, \xi) = \begin{cases} \exp\left\{-[1 + \xi \left(\frac{x - \mu}{\sigma}\right)]^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left(-\frac{x - \mu}{\sigma}\right)\right\}, & \xi = 0 \end{cases} \tag{2.1}$$

and have a probability distribution function (pdf) like the following equation:

$$f(\mu, \sigma, \xi) = \begin{cases} \frac{\xi^{-1}}{\sigma} \exp\left\{-[1 + \xi \left(\frac{x - \mu}{\sigma}\right)]^{-\frac{1}{\xi}}\right\} \exp\left\{-[1 + \xi \left(\frac{x - \mu}{\sigma}\right)]^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \frac{1}{\sigma} \exp\left(-\frac{x - \mu}{\sigma}\right) \exp\left(-\exp\left(-\frac{x - \mu}{\sigma}\right)\right), & \xi = 0 \end{cases} \tag{2.2}$$

where $x$ is the extreme value obtained from the BM method. GEV has three distribution parameters, that known as location parameters/μ, scale parameters/σ, and shape parameters/ξ. Form of the parameters follow 3 distributions, namely Rev. Weibull ($\xi < 0$), Gumbel ($\xi = 0$), and Frechet ($\xi > 0$) [6]. The difference of the distribution can be found in the tip of the three shape parameters which makes the distribution difficult to observe so the GEV distribution serves to combine the three by following the GEV distribution form.

2.2 Maximum Likelihood Estimation (MLE)
The GEV estimation parameter can be done by using the Maximum Likelihood Estimation (MLE) method. The following are parameter estimates for the parameters $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ with MLE using pdf from GEV like equation (2.2). Then the likelihood function with $\xi \neq 0$ is:

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{n} \left\{\frac{1}{\sigma} \exp\left\{-[1 + \xi \left(\frac{X_i - \mu}{\sigma}\right)]^{-\frac{1}{\xi}}\right\} \exp\left\{-[1 + \xi \left(\frac{X_i - \mu}{\sigma}\right)]^{-\frac{1}{\xi}}\right\}\right\}$$

$$\tag{2.3}$$
From the results of the first derivative (2.3), if the form of the likelihood function is not explicit, a numerical approach is needed, namely the Newton Raphson method.

2.3 Dependencies
In statistical theory, modelling EVT in a location is considered insufficient to deal with various extreme events because extreme events involve many adjacent locations that are influential factors. So the extreme events observed in two or more different locations require an approach that is the max-stable approach and the copula approach [7].

2.3.1 Max-Stable Approach
This method used to stabilize data by transforming extreme value distribution (GEV) into the Frechet distribution. This is needed because the three types of GEV distribution, the most heavy-tailed distribution (\( \xi > 0 \)) is the Frechet distribution. Therefore, the process of stabilizing data from extreme value distribution into the Frechet distribution be required that known as the max-stable process. The transformation process of cdf GEV into Frechet pdf using the following equation:

\[
z(x) = \left\{ 1 + \frac{\xi(x - \bar{x})}{\bar{\sigma}} \right\}^{\frac{1}{\xi}}
\]  

(2.4)

where \( z \) is a max-stable process with a Frechet distribution, \( \hat{\mu}, \hat{\sigma}, \hat{\xi} \) are parameter estimates from GEV, and \( x_{ij} \) is the extreme value of the first observation at the observation station \( j \). Based on equation (2.5), there are several models that formed of max-stable processes. One model that usually used is the Smith model. This model is considered the simplest model by involving normal distribution [7]. This model has the following PDF:

\[
f(z_j(x), z_{j+1}(x)) = \exp\left\{ \frac{1}{z_{j+1}} \Phi \left( \frac{h}{2} + \frac{1}{z_j} \log \left[ \frac{z_{j+1}}{z_j} \right] \right) - \frac{1}{z_{j+1}} \Phi \left( \frac{h}{2} + \frac{1}{z_j} \log \left[ \frac{z_j}{z_{j+1}} \right] \right) \right\}
\]  

(2.5)

where \( h \) is the euclid distance from 2 locations \( j \) and \( j + 1 \), \( z \) is the max-stable process with equation (2.5), and \( \Phi \) is a cdf from the normal distribution. To get the spatial dependencies, the trend surface model can be used. This model is able to explain dependencies between locations using the coordinates of the location of rainfall measurements, so that the trend surface model follows the following equation [2]:

\[
\begin{align*}
\mu_j &= \beta_{1,0} + \beta_{1,1} lon_j + \beta_{1,2} lat_j \\
\sigma_j &= \beta_{2,0} + \beta_{2,1} lon_j + \beta_{2,2} lat_j \\
\xi_j &= \beta_{3,0}
\end{align*}
\]  

(2.6)

where \( lon_j \) is the longitudinal coordinate and \( lat_j \) is the latitude coordinate in the \( j \)-rainfall station. So to get the parameter estimation in a number of observation stations \( j \), the maximum likelihood estimation method is used. So the parameter estimation of \( \hat{\beta} \) is:

\[
\hat{\beta} = \arg \max L(\beta, z)
\]

However, the parameter estimates for max-stable approach are considered if applied in many locations. This means that the more locations used, the greater error value generated by the AIC method [8]. So that another approach is carried out, namely using the Copula approach which can connect the number of locations [9].

2.3.2 Copula Approach
Copula was introduced by Abe Sklar in 1959 through the sklar theorem. According to the sklar theorem, copula is a function that connects multivariate distribution functions with marginal distribution [9]. Copula can be explore and characterize dependency structures between random variables through marginal distribution function. Copula is divided into two types of families, namely elliptical copula and archimedian copula. For the case of Spatial Extreme the copula model that can be used is copula elliptical. Copula which is included in the elliptical copula is Copula Gaussian and Student’s t-copula [2]. In this study Gaussian copula was used because this copula model is the most popular model used and is
a simple model. In general, copula can be used universally with uniform standard functions. So the combined distribution function of the random variable x that follows the uniform standard distribution is:

\[ C(u_1, ..., u_n) = F(F'(u_1), ..., F'(u_n)) \]

where F is a combined pdf of random variable x. So that it can be defined that:

\[ f(x_1, ..., x_n) = c(F_1(x_1), ..., F_n(x_n)) \]

In this study, extreme value copula is used as a spatial approach. Extreme value copula is more suitable for heavy-tailed data, when the parameter \( \xi > 0 \). From the three types of GEV distribution, namely reversed weibull, gumbel, and Frechet, the distribution that is the most heavytail is the Frechet distribution. So the parameter estimation results from GEV will follow the Frechet distribution where the parameter \( \xi \) for the Frechet distribution is \( \xi > 0 \) [10]. In the Copula Gaussian for the extreme spatial case the transformation process uses the GEV cdf with the transformation equation defined in the following equation:

\[ u_i = F_j(x_{ij}) \]  (2.7)

where \( F_j \) is cdf GEV and \( x_{ij} \) is j-th and i observation data. So copula cdf follows the following equation:

\[ C(u_1, ..., u_n) = \Phi(f'(u_1), ..., f'(u_n)) \]  (2.8)

where \( \Phi \) is cdf from the normal distribution because the copula used is copula gaussian. So according to the Sklar theorem, copula has a opportunity joint function as follows:

\[ f(x_{11}, x_{12}, ..., x_{ij}, ..., x_{i1}, x_{i2}, ..., x_{ij}, u_2, ..., u_j) \]  (2.9)

where f is pdf GEV because the distribution used is the GEV distribution. After transformation into copula, the spatial GEV estimation parameter of copula will be performed using MLE. So the result will be obtaining from the parameter estimation copula using trend surface models like equation (2.9) [11]. So it is necessary to estimate the parameter \( \hat{\beta} \) using the joint opportunity function with copula. So the form of the log-likelihood from the pdf above is:

\[ \ln(\hat{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \ln f(x_{ij}, \epsilon(u_1, ..., u_n)) \]

where \( x_{ij} \) is the i-extreme value at the jth observation station.

2.4 Return Level

To examine extreme events, determine the return value is also an important thing. Return value is the maximum value that is expected to occur in the coming period. The concept used in the return value is to predict rare events that will occur such as extreme rainfall that has a negative impact. Return Level is also an application of GEV distribution. From the cdf on GEV, the probability of extreme rainfall is 1-p, where p is the return period with \( k \) is the number of blocks formed so that [12]:

\[ \tilde{R}_j = \hat{\mu}_j - \frac{\tilde{\xi}_j}{\hat{\beta}_j} \left[ 1 - \left( -\ln(1-p) \right)^{-\tilde{\xi}_j} \right] \]  (2.10)

Which means that the return value from extreme events on the jth observation station can be determined using the equation. Where \( \hat{\mu}_j \) is the estimation of the jth location parameter, \( \hat{\sigma}_j \) is the estimation of the jth scale parameter, \( \tilde{\xi}_j \) is the estimation of the jth shape parameter, and p is return period.

2.5 Akaike’s Information Criterion (AIC)

Akaike’s Information Criterion (AIC) is the best model selection method. So this method is used to determine or select the best model that will be produced from this study. In the certain contexts, choosing a simple model is better than choosing a complex model. According to Ligas and Banasick [13] AIC is defined by the following equation:

\[ AIC = -2 \log \log L(\beta) + 2q \]  (2.11)

Where \( \log L(\beta) \) is the log-likelihood function of each approach and q is the number of parameters estimated. A lower AIC value indicates the best model that can be used.
2.6 Extreme Rainfall
Rainfall is the amount of water that falls on a flat land surface during a certain period measured in millimeters (mm). Extreme rainfall has parameters with rainfall intensity > 100 millimeters per day, while rainfall with intensities > 50 millimeters per day is included in rainfall with heavy intensity. (BMKG, 2016). Rainfall is data that classified into spatial data. So, spatial extreme rainfall is extreme rainfall that occurs in various locations and has spatial data elements such as coordinates and the others [14].

3. Result and Discussion
In this study, we used secondary data that obtained directly from the website of the Meteorology and Geophysics Agency (BMKG), namely dataonline.bmkg.go.id. The data used is rainfall data in East Java which consists 9 of rainfall observation stations starting in 2009 until 2018.

The first step is make 4 periods (blocks) of rainfall in one year which in one block contains 3 months rainfall data based on the moonsun rainfall pattern, namely the December-January-February block (DJF), March-April-May block (MAM), June-July-August block (JJA), and September-October-November block (SON). After forming many blocks, the maximum value of each block will be taken, called the Block Maxima (BM) method. So in the intervals from 2009-2018, 40 blocks will be formed. From the results of the block maxima method, a correlation between the locations obtained will be checked. So it can be concluded that all locations have a correlation or interconnection between locations with each other. Because the distribution of extreme values taken using the BM method will follow the GEV distribution, GEV parameter estimation will be carried out univariately using maximum likelihood estimation with 3 parameters, namely location parameters, scale parameters, and form parameters.

Because the data used contains many locations, it is indicated that there are spatial elements that affect extreme rainfall that occurs. So the approach that can be used is the max-stable approach and the copula approach [15]. The first step to calculate dependencies using the max-stable approach is to transform the extreme value distribution into the Frechet distribution. This process is called the max-stable process. This transformation uses equation (2.8). To obtain spatial parameter estimates with this approach, the Smith model is used, followed by a trend surface model. As for the copula approach, it uses the copula unit transformation so that it forms a gaussian copula model. From the copula gaussian model, parameter estimates can be obtained which are also followed by the trend surface model. The parameter estimates used by both approaches are MLE.

After obtaining spatial parameter estimates, the rainfall prediction values obtained from each approach are used. So that in the following tables of return value with a 2-years return period at each station can be shown in the following table:

Table 1. Return level of max-stable approach

| Station   | 2019-2020 | 2021-2022 | 2023-2024 |
|-----------|-----------|-----------|-----------|
| Banyuwangi| 274.55    | 338.84    | 368.78    |
| Juanda    | 527.54    | 660.56    | 722.5     |
| Kalianget | 327.55    | 410.37    | 448.94    |
| Karangkates| 590     | 735.93    | 803.89    |
| Malang    | 563.13    | 703.17    | 768.39    |
| Perak 1   | 535.88    | 671.73    | 734.98    |
| Sangkapura| 534      | 675.07    | 740.77    |
| Sawahan   | 699.44    | 875.34    | 957.25    |
| Tretes    | 666.08    | 693.87    | 758.5     |
Table 2. Return level of copula approach

| Station       | Year      |        |        |        |
|---------------|-----------|--------|--------|--------|
|               | 2019-2020 | 2021-2022 | 2023-2024 |        |
| Banyuwangi    | 293.12    | 379.9  | 423.19 |        |
| Juanda        | 562.06    | 737.87 | 825.59 |        |
| Kalianget     | 334.79    | 438.16 | 489.73 |        |
| Karangkates   | 642.81    | 842.64 | 942.33 |        |
| Malang        | 609.26    | 798.92 | 893.54 |        |
| Perak 1       | 568.95    | 747.43 | 836.47 |        |
| Sangkapura    | 546.53    | 721.47 | 808.75 |        |
| Sawahan       | 758.33    | 996.55 | 1115.39 |       |
| Tretes        | 597.41    | 783.7  | 876.75 |        |

From table 1 and 2, it can be concluded that the highest intensity rainfall occurs around the Sawahan rainfall observation station both with the max-stable approach and the copula approach respectively. Whereas to determine the best results from the two approaches used, the Akaike Information Criterion (AIC) method is used. Based on this method, according to Ligas and Banasick [13] an approach that has the lowest AIC value is the best approach that can be used. So that the results of the AIC method are as follows:

Table 3. AIC value

| Approach           | AIC Value |
|--------------------|-----------|
| Max-stable approach| 37764.95  |
| Copula Gaussian    | 4549.22   |

Therefore, from Table 3 it can be said that the model using copula Gaussian is the best model compared with max-stable approach as it provides the smallest AIC values [2].

4. Conclusion

The return level for extreme rainfall events in East Java with two approach give the result that the highest intensity of rainfall will occur in the same location which is around the Sawahan rainfall observation station located in Nganjuk district. But it has different rainfall intensity. The best model produced based on the AIC is the copula approach with an AIC value of 4549.22.

References

[1] C. E. R. Cooley, “A Survey of spatial extremes:Measuring Spatial Dependence and Modelling Spatial Effect,” *Stat. J.*, 2012.
[2] S. P. M. R. Davison A.C, “Statistical modelling of spatial extremes,” *Stat. Sci.*, 2012.
[3] H. T. S. A. Oktaviarina, A. Sofro, “Generalized Extreme Value for Climate Change Analysis,” *Atl. Highlights Eng.*, 2018.
[4] L. F. E. McNeil, “Modelling dependence with copulas and applications to risk management,” *Elsevier*, 2003.
[5] Gili M. Kellezi, “An application of extreme value theory for measuring financial risk,” *Comput. Econ.*, 2006.
[6] Schlather, “Models for stationary max-stable random,” 2002.
[7] R. A. Davis, “Statistical inference for max-stable processes in space and time,” *J. R. Stat. Soc.*, 2013.
[8] F. Ramadhan, “Extreme Value Theory (EVT) Application On Estimating The Distribution Of Maxima,” in *Conference Proceedings*, 2017.
[9] Nelsen, “An Introduction to Copula,” *Springer Ser. Stat.*, 2006.
[10] M. Ribatet, “Spatial Extremes Package.” 2018.
[11] J. Tawn, “Modelling spatial extreme events with environmental applications,” Spat. Stat., 2017.
[12] A. Rudvik, “Dependence Structures in Stable Mixture Models with an Application to Extreme Precipitation,” 2012.
[13] Ligas M. and Banasik P., “Local height transformation through polynominal regression,” J. Geod. Cartogr., 2012.
[14] L. D. Haan, “A spectral representation for max-stable processes,” Ann. Probab., 1984.
[15] WeiB, “Copula Parameter Estimation: Numerical Considerations and Implications for Risk Management,” J. Risk, 2010.