ICIC Express Letters
Part B: Applications
Volume **12**, Number **4**, April **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **403–410**

EXTREME RAINFALL PREDICTION BASED ON BLOCK MAXIMA DATA REPRESENTATION USING VAR-RBFN MODEL

Hasbi Yasin*, Suparti, Puspita Kartikasari, Budi Warsito and Rukun Santoso

Department of Statistics
Faculty of Sciences and Mathematics
Diponegoro University
Prof. Soedarto, S.H. Road, Tembalang, Semarang, Central Java 50275, Indonesia
*Corresponding author: hasbiyasin@live.undip.ac.id
puspitakartikasari@live.undip.ac.id; {suparti702; budiwarsito1975; rukusantoso25}@gmail.com

Received September 2020; accepted December 2020

Abstract. *This study aims to compile a computational application to predict extreme rainfall. In this study, daily rainfall data was transformed into a three-monthly block data using the block maxima methods. Based on these transformations, extreme rainfall is obtained which is used as a basis for predicting extreme rainfall that may occur. Rainfall data were collected from several rain observation stations in the Semarang city, Central Java. Therefore, to estimate the extremes of several locations simultaneously used the multivariate time series approach. One of these methods is called Vector Autoregressive (VAR) models. Because of the limitations of the VAR model in nonlinearity, the parameters of VAR will be estimated by Neural Network (NN) especially Radial Basis Function Network (RBFN) and called the VAR-RBFN model. In this study, the optimized parameters were variable dependent lag and the number of neurons in the hidden layer. The results showed that model VAR(4)-RBFN(29) is the most appropriate model for predicting extreme rainfall in the city of Semarang based on the smallest SMAPE value. This model has an average error rate of 5.194%.*
Keywords: Block maxima, Rainfall, RBFN, Vector autoregressive

1. **Introduction.** High rainfall will cause some areas in Indonesia to be hit by floods. According to the daily rainfall data from the Climatology Station of Semarang city, Central Java, Indonesia, currently (as 2019) the maximum daily rainfall ranges have reached 80-100 mm per hour [1,2]. Semarang city is the most important city in Central Java province. It is also an economic center that is connected to other regions in Central Java. Extreme rainfall can trigger natural disasters that can disrupt economic access, government activities, food security, and hamper industrial sustainability. In rainfall data, nonlinearity modeling is also required to obtain the best models. Some studies of the prediction of extreme rainfall have been conducted especially in Indonesia. Among them is a prediction of extreme rainfall in the Ngawi District using spatial extreme value based on max stable process [3]. One method in determining the extreme rainfall is the Block Maxima (BM) method. In this study, rainfall data is presented in the block maxima with three months of Dec-Jan-Feb (DJF), Mar-Apr-May (MAM), Jun-Jul-Aug (JJA), and Sep-Oct-Nov (SON) [4].

Recently, many prediction methods are proposed to forecast the chaotic time series data such as the fuzzy time series model [5] and time series regression model [6]. The time series regression method is not only applied for one time series variable but also two or more vectors simultaneously (multivariate cases). Vector Autoregressive (VAR) is one of some models that can be applied to predicting multivariate data. Multivariate time

DOI: 10.24507/icicelb.12.04.403

series data can consist of different vector variables observed in a certain period, or it can consist of the same variables and in the same observation period but observed at several locations. VAR is the linear model that can be applied not only on stationary data [7]. In the VAR model, it is required to specify the predictor variable using the lag of the variables used. Some problems in the VAR model are stationarity of data, linearity, and normality of the error model [8]. These problems can be solved using the NN approach, so this model is known as the VAR-NN model. This model is using the predictor variables in the VAR model as the neuron in the input layer [7,8].

Many studies show that NN provides high accuracy in time series analysis. Some of them are rainfall prediction using VAR-NN in Malang, Indonesia [11], prediction airline passenger using VAR-NN [7], and VAR-NN to forecast the oil production [6]. In this study, the Radial Basis Function Network (RBFN) is used to estimate the model parameters that were conducted using a vector autoregressive approach and namely the VAR-RBFN model. RBFN is a type of NN that is widely applied for predicting time series data [12-14]. This model is used to predict the extreme rainfall in Semarang city using the block maxima data approach. The problem that arises in VAR-RBFN modeling is how to select the lag of the dependent variable and the optimal number of neurons in the hidden layer. Therefore, in this study, a user interface of the VAR-RBFN model will be developed to assist in the selection of VAR-RBFN parameters [15]. The best model is selected based on the smallest Symmetric Mean Absolute Percentage Error (SMAPE) value [16,17].

2. **Materials and Method.**

2.1. **Block maxima.** BM is a method that is used to identify extreme values by capturing the maximum values in a certain period or block. In the Extreme Value Theory (EVT), the block maxima approach is obtained by dividing the observation period into non-overlapping periods of the same size and limiting attention to maximum observations in each period [18]. The maximum value in each block is a sample of the extreme values in that period. An illustration of how to collect extreme data using the block maxima method in this study can be seen in Figure 1.

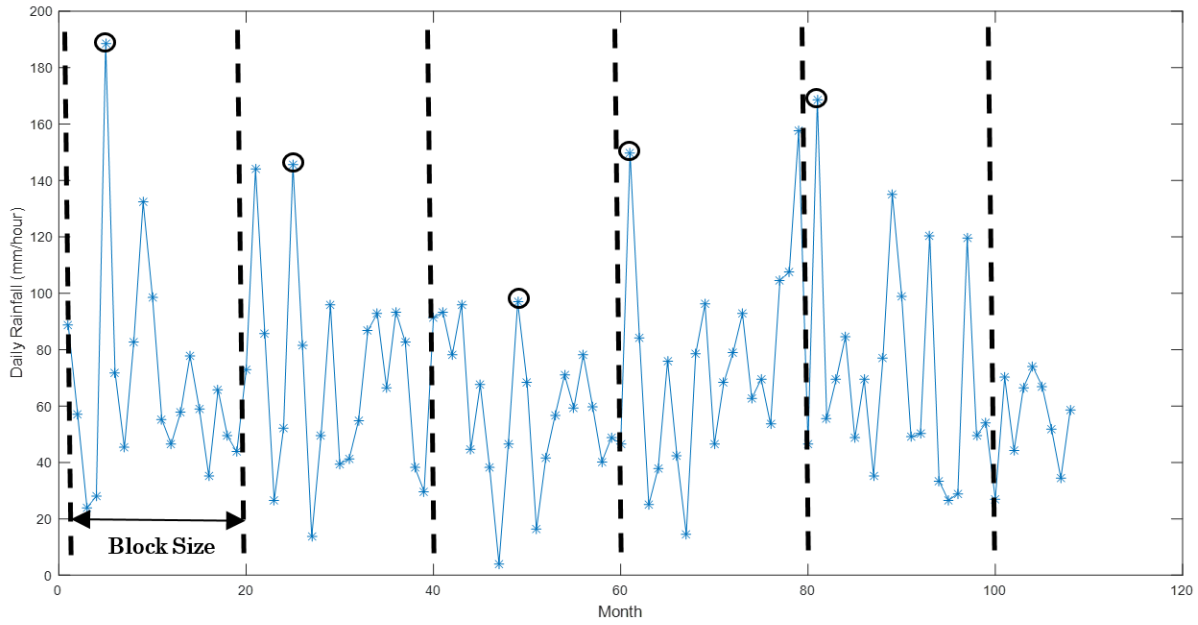

FIGURE 1. Block maxima illustration

2.2. **Vector autoregressive model.** The Vector Autoregressive (VAR) model is one of statistical methods that are applied to modeling the time series data using the past data of the variable by the multivariate approach. The Vector Autoregressive (VAR) model is a combination of several Autoregressive (AR) models that form a vector that affects each other [7]. The $VAR(1)$ model is a vector autoregressive model with order 1. The $VAR(1)$ means that the independent variable of the model is only one lag dependent variable [19]. Let $\boldsymbol{Z}(t)$ be the matrix of dependent series in time (t), p is the number of lags, and $\boldsymbol{\phi}$ is the coefficient of VAR model, and then $VAR(p)$ model with three series of variables is

$$\boldsymbol{Z}(t) = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{Z}(t-i) + \boldsymbol{\varepsilon}(t) \tag{1}$$

$$\boldsymbol{Z}(t) = \begin{bmatrix} Z_1(t) \\ Z_2(t) \\ Z_3(t) \end{bmatrix}, \boldsymbol{Z}(t-i) = \begin{bmatrix} Z_1(t-i) \\ Z_2(t-i) \\ Z_3(t-i) \end{bmatrix}, \boldsymbol{\phi}_0 = \begin{bmatrix} \phi_{10} \\ \phi_{20} \\ \phi_{30} \end{bmatrix}, \boldsymbol{\phi}_i = \begin{bmatrix} \phi_{11}^{(i)} & \phi_{12}^{(i)} & \phi_{13}^{(i)} \\ \phi_{21}^{(i)} & \phi_{22}^{(i)} & \phi_{23}^{(i)} \\ \phi_{31}^{(i)} & \phi_{32}^{(i)} & \phi_{33}^{(i)} \end{bmatrix},$$

$$\boldsymbol{\varepsilon}(t) = \begin{bmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \varepsilon_3(t) \end{bmatrix}.$$

2.3. **Radial basis function network.** RBFN is one type of NN commonly used to solve forecasting problems [14]. In MATLAB, RBFN is divided into two functions, namely *newrbe* and *newrb*. Both have almost the same characteristics in the learning process, and the difference is that the *newrbe* has the number of neurons as much as the number of inputs that are formed, while in *newrb* each iteration will be formed 1 neuron. The radial basis function network is designed to form a mapping from input variables to hidden layer units and mapping from hidden layer to output. Therefore, there are three layers on the radial base function structure: the input layer, hidden layer, and outputs layers. The input layer contains points (nodes) that are composed as many predictor variables. The hidden layer consists of hidden units that activate using a radial basis function. And the last layer consists of a single linear unit that shows the predicted value. Suppose that the set of N training data points has P input vectors, and each input vector $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iP}\}$; $i = 1, 2, \ldots, N$ with one output vector $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ then the RBFN equation with J neuron in hidden layer can be expressed as:

$$y_i = w_0 + \sum_{j=1}^{J} w_j \phi_{ij}(\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|) \tag{2}$$

and in matrix form, Equation (2) can be written as follows:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w} \tag{3}$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $\boldsymbol{\Phi} = \begin{bmatrix} 1 & \phi_{11} & \phi_{12} & \cdots & \phi_{1J} \\ 1 & \phi_{21} & \phi_{22} & \cdots & \phi_{2J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_{N1} & \phi_{N2} & \cdots & \phi_{NJ} \end{bmatrix}$, and $\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_J \end{bmatrix}$.

In the formula w_j is the connection weight from the jth hidden node to the output, ϕ is the radial basis activation function, $\|.\|$ represents the Euclidean distance, $\boldsymbol{\mu}_j = \{\mu_{j1}, \mu_{j2}, \ldots, \mu_{jP}\}$; $j = 1, 2, \ldots, J$ is a center vector as neuron in the jth hidden layer. If Gaussian radial basis function is used and the spread (σ) is set to the equal fixed value, then the output function of the network is written as follows [20]:

$$y_i = w_0 + \sum_{j=1}^{J} w_j \exp\left(-\frac{\ln(0.5)\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2}{\sigma^2}\right) \tag{4}$$

The optimal approach for the RBF parameters is to set all their spreads to be fixed and equal at an appropriate size for the distribution of data points and to have their centers fixed at J points selected at minimum error from the N data points. In particular, we can use normalized RBFs centered at $\{\boldsymbol{\mu}_j\} \subset \{\boldsymbol{x}_i\}$ and the spread (σ) related to the average Euclidean distance between the chosen centers and the training data [21].

2.4. **VAR-RBFN.** RBFN is one of the most common models that are used in time series forecasting. In general, to predict more than one variable simultaneously used the multivariate approach. Therefore, the VAR-RBFN is proposed which is derived from the VAR model. This network architecture resembles the classical RBFN but there is more than one neuron in the output layers [19]. In VAR-RBFN, the lag dependent variables from the previous time are being the neuron in the input layer. For example, 3 variables will be predicted using the 2 lag dependent variable, and then the input neuron will consist of 6 neurons. In matrix form, the *VAR(p)-RBFN(J)* model of 3 vector time series data with $p = 2$ lag dependent variable in the input vector and J neuron in hidden layer can be written as follows:

$$\boldsymbol{Z}(t) = \boldsymbol{\Phi}\boldsymbol{W} \tag{5}$$

where

$$\boldsymbol{Z}(t) = \begin{bmatrix} \boldsymbol{z}_1(t) & \boldsymbol{z}_2(t) & \boldsymbol{z}_3(t) \end{bmatrix} = \begin{bmatrix} z_{11}(t) & z_{12}(t) & z_{13}(t) \\ z_{21}(t) & z_{22}(t) & z_{23}(t) \\ \vdots & \vdots & \vdots \\ z_{N1}(t) & z_{N2}(t) & z_{N3}(t) \end{bmatrix},$$

$$\boldsymbol{W} = \begin{bmatrix} w_{01} & w_{02} & w_{03} \\ w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ \vdots & \vdots & \vdots \\ w_{J1} & w_{J2} & w_{J3} \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} 1 & \phi_{11} & \phi_{12} & \cdots & \phi_{1J} \\ 1 & \phi_{21} & \phi_{22} & \cdots & \phi_{2J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_{N1} & \phi_{N2} & \cdots & \phi_{NJ} \end{bmatrix},$$

$$\phi_{ij} = \exp\left(-\frac{\ln(0.5)\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2}{\sigma^2}\right),$$

$$\boldsymbol{x}_i = \{z_{i1}(t-1), z_{i2}(t-1), z_{i3}(t-1), z_{i1}(t-2), z_{i2}(t-2), z_{i3}(t-2)\}; \quad i = 1, 2, \ldots, N,$$

$$\boldsymbol{\mu}_j = \{z_{j1}(t-1), z_{j2}(t-1), z_{j3}(t-1), z_{j1}(t-2), z_{j2}(t-2), z_{j3}(t-2)\}; \quad j = 1, 2, \ldots, N.$$

2.5. **Study area and data used.** In this research, we use the daily rainfall data from the Climatology Station of Semarang city, Central Java, Indonesia. The data used was collected from three Rainfall Observation Stations in Semarang city from 1990 to the 2019 period. The rain monitoring stations in Semarang city are: *Stasiun Klimatologi Semarang*, *Stasiun Meteorologi Ahmad Yani*, and *Stasiun Meteorologi Maritim Tanjung Mas*. Figure 2 shows a time series plot (a) and a correlation plot (b) of extreme rainfall data at the three stations analyzed in this study. This figure shows that the extreme rainfall in Semarang has reached 276 mm/hour in a day which caused severe flooding in 1993 [22,23]. Based on this figure, it can be seen that the extreme rainfall patterns of the three stations are very similar, as evidenced by the significant correlation value at the 1% level (see legend of Figure 2). Therefore, the data in this study are multivariate time series data, so the model used to predict rainfall at these three stations would be better if the VAR-RBFN method was used.

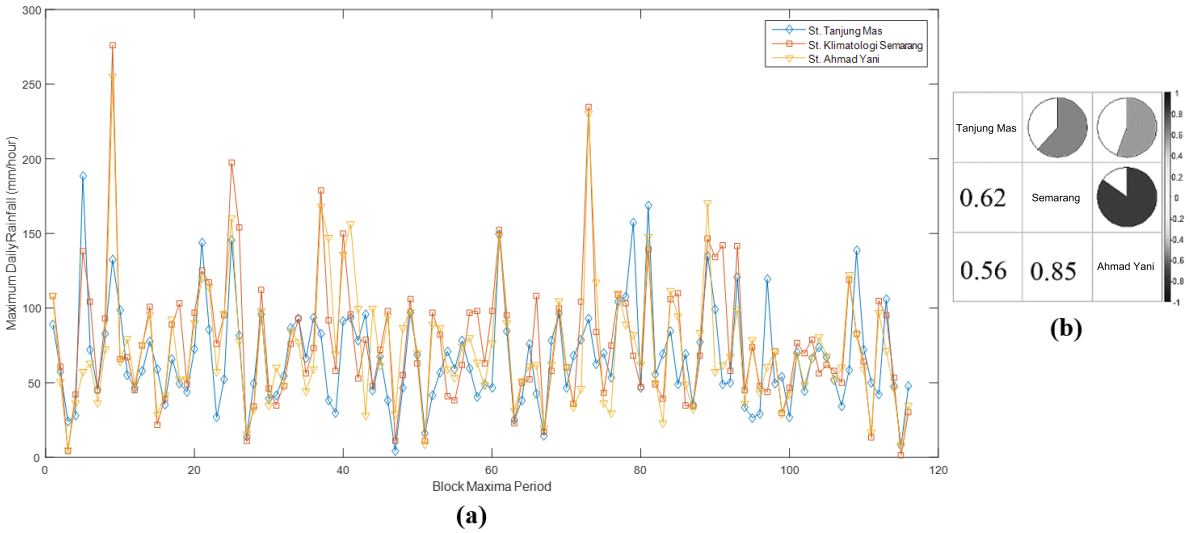

FIGURE 2. Time series plot (a) and correlation plot (b) of rainfall in block maxima data

3. **Main Results.** In this research, to estimate the VAR-RBFN model parameter, we develop a graphical interface application (see Figure 3). Only by entering the percentage of training data set and the lag dependent variable in *VAR(p)-RBFN(J)*, SMAPE values for training and testing data set will be obtained based on the weights matrix which is estimated by Algorithm 1. The rainfall data is divided into two parts, the training

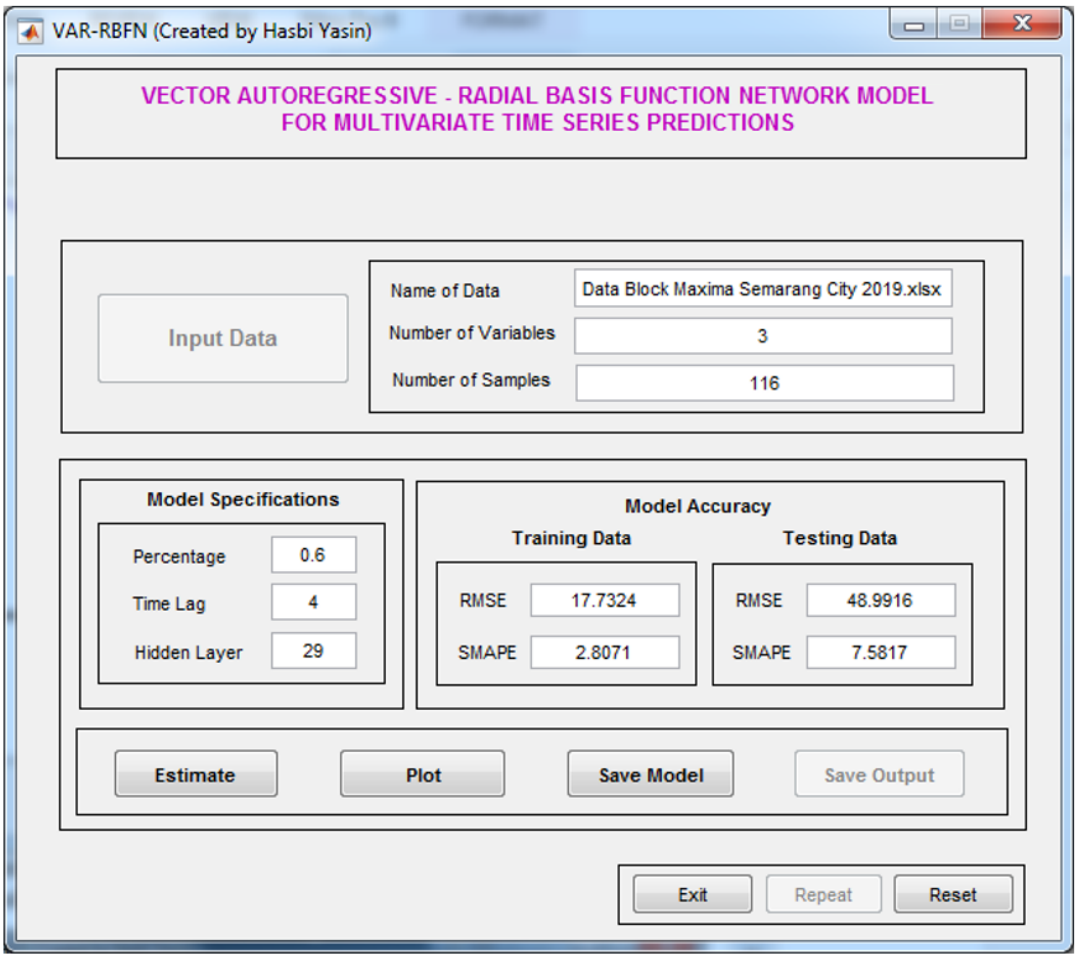

FIGURE 3. Graphical interface of VAR-RBFN model

Algorithm 1. Parameter estimation *VAR(p)-RBFN(J)* using least square method

Input: Matrix of p lag dependent variable as neuron input $\boldsymbol{Z}(t-p)$
Number of neurons in hidden layer (J)

Output:
1. Preprocess data by taking standard normal transformations.
2. Calculate Euclidean distance between each input vector with the centers: $\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|$
3. Calculate the spread of radial basis functions (σ), using average of the Euclidean distance in step 1.
4. Calculate radial basis activation function: $\phi_{ij} = \exp\left(-\frac{\ln(0.5)\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2}{\sigma^2}\right)$
5. Calculate the weight matrix $\boldsymbol{W}$ using least square method.
$$\widehat{\boldsymbol{W}}(t) = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{Z}(t)$$
6. Calculate the fitted value by simulation of the network
$$\widehat{\boldsymbol{Z}}(t) = \boldsymbol{\Phi}\widehat{\boldsymbol{W}}$$
7. Post processing data
8. Model evaluation by calculating the SMAPE

Table 1. Models comparison based on SMAPE

Experiment	Lag dependent variable	The best neuron in hidden layer	SMAPE (%)		
			Training	Testing	Average
90:10	1	29	5.744	8.025	6.884
	2	12	3.731	7.141	5.436
	3	29	4.049	7.382	5.715
	4	30	3.909	7.214	5.562
80:20	1	4	6.546	5.883	6.215
	2	12	5.703	6.424	6.063
	3	3	6.639	6.247	6.443
	4	23	4.421	6.045	5.233
70:30	1	4	6.186	6.551	6.368
	2	7	5.631	5.908	5.769
	3	23	3.811	7.862	5.836
	4	24	3.475	8.124	5.800
60:40	1	10	5.650	7.038	6.344
	2	4	4.891	6.948	5.920
	3	30	2.832	8.537	5.685
	4	**29**	**2.807**	**7.582**	**5.194**

and testing sets using four experiments by percentage of overall data, which are (90:10), (80:20), (70:30), and (60:40) respectively. Then each data set will be simulated using 1 to 4 lag dependent variables and 1 to 30 neurons in the hidden layer to find the best VAR-RBFN model based on the smallest SMAPE value. The complete simulation results are presented in Table 1. Based on the smallest of average SMAPE value from training and testing data set, the best model to describe this data is the *VAR(4)-RBFN(29)* model. This means that to predict extreme daily rainfall in the Semarang city, we need the information from the 4 previous lag dependent variables (one previous year) and used as many as 29 observations from the training data set. This model has an average SMAPE of 5.194%. It means that this model can predict the extreme rainfall data in Semarang city with high accuracy [24]. By using the best model, the forecast value of extreme rainfall for the next five years at each station can be seen in Figure 4. The forecast results show that the maximum daily rainfall will be around 175 mm/hour.

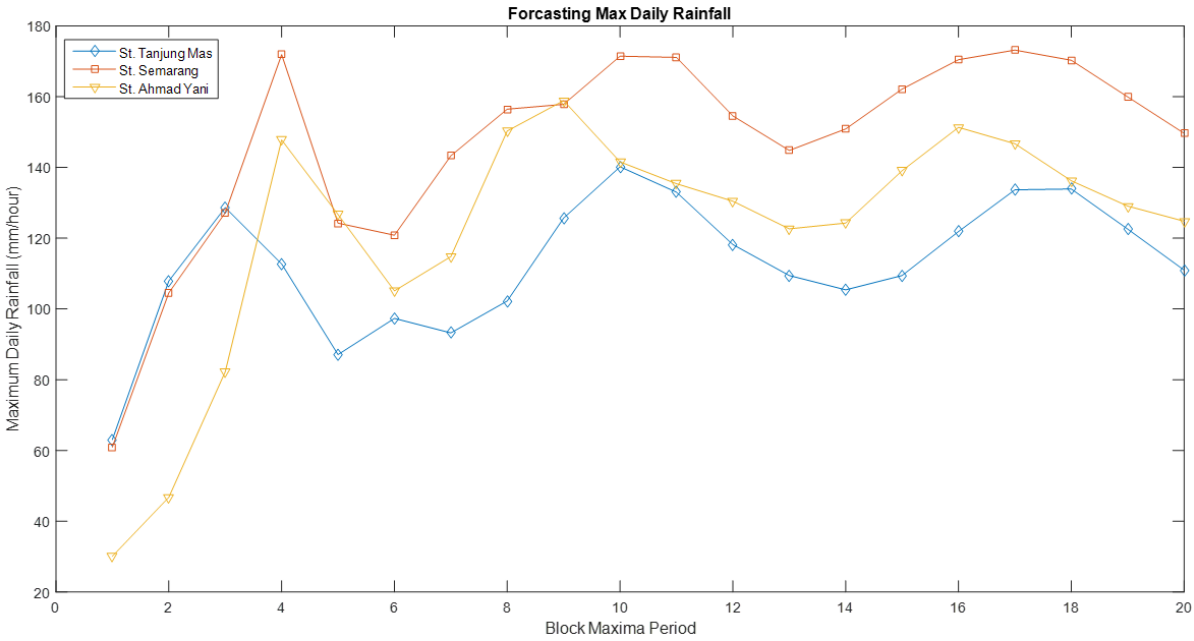

FIGURE 4. Time series plot of forecast value of extreme daily rainfall in the Semarang city for the next 5 years

4. **Conclusion.** In this research, we develop a hybrid method from VAR and RBFN to model an extreme daily rainfall in Semarang city as a multivariate time series approach called VAR-RBFN. The best model in this research is *VAR(4)-RBFN(29)*. The results show that the accuracy of the model is very high with the average SMAPE value of training and testing data set is 5.194%. Therefore, we recommend *VAR(4)-RBFN(29)* using 4 lag of rainfall data to predict the extreme rainfall of the three stations in the Semarang city. Further study is needed to optimize the weights of the VAR-RBFN model using the metaheuristic algorithm to find the high accuracy of the multivariate time series model.

Acknowledgments. This research was fully supported and funded by "Riset Madya" Research Grant 2020 under the Research Assignment Letter No: 2024/UN7.5.8/PP/2020. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] STAKLIM SEMARANG, *Climatology Station of Semarang*, https://www.iklimjateng.info/, 2019.
[2] BMKG, *The ONLINE DATA – DATABASE CENTER – BMKG Application*, http://dataonline.bmkg.go.id/home, 2019.
[3] A. R. Hakim, S. Sutikno and D. D. Prastyo, Spatial extreme value modeling using max-stable processes approach (Case study: Rainfall intensity in Ngawi), *Proc. of the 3rd International Conference on Research, Implementation and Education of Mathematics and Science*, pp.16-17, 2016.
[4] H. Yasin, A. R. Hakim, B. Warsito and R. Santoso, Extreme rainfall prediction using spatial extreme value by max stable process (MSP) Smith model approach, *J. Phys. Conf. Ser.*, vol.1217, no.1, 2019.
[5] M. N. Alemu, A fuzzy model for chaotic time series prediction, *International Journal of Innovative Computing, Information and Control*, vol.14, no.5, pp.1767-1786, 2018.
[6] S. Suhartono, D. D. Prastyo, H. Kuswanto and M. H. Lee, Comparison between VAR, GSTAR, FFNN-VAR and FFNN-GSTAR models for forecasting oil production, *Matematika*, vol.34, no.1, pp.103-111, 2018.
[7] D. U. Wutsqa, Subanar, S. Guritno and Z. Sujuti, Forecasting performance of VAR-NN and VARMA models, *Proc. of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications*, pp.194-200, 2006.

[8] A. D. Aydin and S. C. Cavdar, Comparison of prediction performances of artificial neural network (ANN) and vector autoregressive (VAR) models by using the macroeconomic variables of gold prices, Borsa Istanbul (BIST) 100 index and US dollar-Turkish Lira (USD/TRY) exchange rates, *Procedia Econ. Financ.*, vol.30, no.15, pp.3-14, 2015.
[9] D. U. Wutsqa, The VAR-NN model for multivariate time series forecasting, *J. Mat. Stat.*, vol.8, no.1, pp.35-43, 2008.
[10] H. Yasin, B. Warsito and R. Santoso, Feed forward neural network modeling for rainfall prediction, *E3S Web of Conferences*, vol.73, no.9, 2018.
[11] S. F. Higazi, D. H. Abdel-Hady and S. A. Al-Oulfi, Application of spatial regression models to income poverty ratios in middle delta contiguous counties in Egypt, *Pakistan J. Stat. Oper. Res.*, vol.9, no.1, pp.93-110, 2013.
[12] N. Kanazawa, Radial basis functions neural networks for nonlinear time series analysis and time-varying effects of supply shocks, *J. Macroecon.*, vol.64, no.3, 2020.
[13] A. Tatar, A. Shokrollahi, M. Mesbah, S. Rashid, M. Arabloo and A. Bahadori, Implementing radial basis function networks for modeling CO_2-reservoir oil minimum miscibility pressure, *J. Nat. Gas Sci. Eng.*, vol.15, pp.82-92, 2013.
[14] Z. Wang, Y. Yang, B. Yang and Y. Kang, Optimal sheet metal fixture locating layout by combining radial basis function neural network and bat algorithm, *Adv. Mech. Eng.*, vol.8, no.12, 2016.
[15] H. Yasin, B. Warsito and R. Santoso, Soft computation vector autoregressive neural network (VAR-NN) GUI-based, *E3S Web of Conferences*, 2018.
[16] C. Chen, J. Twycross and J. M. Garibaldi, A new accuracy measure based on bounded relative error for time series forecasting, *PLoS One*, vol.12, no.3, 2017.
[17] V. Kreinovich, H. T. Nguyen and R. Ouncharoen, *How to Estimate Forecasting Quality: A System-Motivated Derivation of Symmetric Mean Absolute Percentage Error (SMAPE) and Other Similar Characteristics*, Departmental Technical Report (CS), 2014.
[18] B. Y. A. Ferreira and L. De Haan, On the block maxima method in extreme value theory: PWM estimators, *Ann. Stat.*, vol.43, no.1, pp.276-298, 2015.
[19] D. A. I. Maruddani and D. Safitri, Vector autoregressive (VAR) for forecasting the stock price of PT Indofood Sukses Makmur Indonesia Tbk, *Junal Mat.*, vol.11, no.1, pp.6-12, 2008.
[20] S. Ding, L. Xu, C. Su and F. Jin, An optimizing method of RBF neural network based on genetic algorithm, *Neural Comput. Appl.*, vol.21, no.2, pp.333-336, 2012.
[21] J. A. Bullinaria, *Radial Basis Function Networks: Algorithms*, 2015.
[22] Y. Yunarto and A. M. Sari, Relocation of flood/rob affected population in Semarang city, *Maj. Ilm. Globe*, vol.19, no.2, p.123, 2017.
[23] R. J. Prakasa, R. Anggoro, A. Kadir and A. Falah, Analysis of the cross section capacity of west Semarang canal floods for flood control planning, *J. Karya Tek. Sipil*, vol.2, no.1, pp.290-308, 2013.
[24] J. J. Montaño Moreno, A. P. Pol, A. S. Abad and B. C. Blasco, The R-MAPE index as a resilient measure of adjustment in forecasting, *Psicothema*, vol.25, no.4, pp.500-506, 2013.