Rainfall Forecasting by Using Autoregressive Integrated Moving Average, Single Input and Multi Input Transfer Function

Ahmad Saikhu[1], Agus Zainal Arifin[2], Chastine Fatichah[3]
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
[1]saikhu@if.its.ac.id, [2]agusza@cs.its.ac.id, [3]chf@cs.its.ac.id

Abstract— **This study aims to compare performance of three methods for forecasting rainfall, i.e. Autoregressive Integrated Moving Average, Single Input and Multi Input Transfer Function. These methods are applied to meteorological data from Indonesian Agency for Meteorology Climatology and Geophysics, Juanda-Surabaya. For the first modeling, using single variable, i.e. rainfall time lags. While for single input and multi input transfer function, using seven local predictor variables, that are minimum, maximum, average temperature, humidity, solar radiation, wind speed and maximum wind speed. The five global predictor variables are added, i.e. Nino1.2, Nino3, Nino3.4, Nino4 and DMI. From the transfer function model, it will be known to any exogenous variables that significantly affect rainfall. Based on the model performance measurement, it is found that the best model is Multi Input Transfer Function with four exogenous predictors, followed by Single Input and the last is univariate Autoregressive Integrated Moving Average. In the time-series modeling, its accuracy will increase when it involves external predictor variables, including rainfall forecasting.**

Keywords— ***Autoregressive Integrated Moving Average, single input, multi input, transfer function, rainfall, predictors.***

I. Introduction

Weather, climate and the season are always familiar in everyday human life. Natural phenomena that affect the lives of many sectors including agriculture, forestry, maritime, fisheries, health, transport, tourism, energy and infrastructure. In extreme conditions, the weather and climate can even cause disasters, including floods, high winds, landslides, droughts and forest fires. Therefore, information analysis and knowledge of weather and climate conditions and patterns of climate change are very important to learn that humans can adapt to dynamic changes avoid or reduce the risk of disasters and mitigation [1].

Rainfall forecasting is a very important problem that arises in many applications, such as agriculture, water resources management, hydrology or facilities maintenance and control. Different previous works that have applied deterministic approach numerical models, statistics and soft computing approach still unable to provide accurate model [2]. An important component in identifying patterns of rainfall is a measurement of weather and climate variables i.e. air temperature, relative humidity, air pressure, water evaporation, wind speed, sea-surface temperature and duration of solar radiation [3, 4].

Statistical techniques which have been used for rainfall forecasting since a past century were basically linear conceptual and statistical models. One of these popular and frequently used models was Autoregressive Inegrated Moving Average (ARIMA) model [5]. Farajzadeh et al. [6] used ARIMA model for monthly rainfall forecasting in northwest of Iran. Nugroho and Simanjuntak [7] used univariate ARMA model for the same data in Regency of Semarang. In recent two decades, ARIMA becomes a valuable method for multivariate phenomenon modeling i.e. Vector Autoregreesion (VAR) rainfall forecast, [8] multi-input transfer function [5] and spatio-temporal [9, 10, 11].

Based on these research outcomes, this research compares performance of three methods for forecasting rainfall, i.e. ARIMA, single-input and multi-input transfer function methods. It aims to evaluate the effectiveness of the forecasting model, whether simply by using the time lags of the variable, i.e. ARIMA or using exogenous variable, single or multi variables as a simultaneously.

In this study, beside local variabels to rainfall forecast we add the global variables and take into account the relationship of time-lags. These local variables are min/max/mean temperature, humidity, wind-speed, maximum wind-speed and solar radiation. The global variables are Nino1.2, Nino3, Nino34, Nino4 and Dipole Mode Index (DMI). These variables, in a number of studies expressed affecting rainfall over an area included at Indonesia region [3, 4, 5].

II. Dataset and Preprocessing

A. Case Study

Local observation data obtained from Indonesian Agency for Meteorology Climatology And Geophysics (BMKG) to weather observation station class 1 In East Java namely Juanda (Coordinates:S 07º22'38";E 112º22'38"). This weather station observes 30 km^2 area in Surabaya and surrounding. Local features was resulted by this station are: daily rainfall minimum/mean/maximum temperature(minT, meanT, maxT), humidity (RH), wind-speed (WS), max wind-speed (max-WS) and solar radiation (SR). Observation daily periods are 36 Years January 1, 1981 until December 31, 2016.

Global variable data (climate indices) was obtained from The Royal Netherlands Meteorological Institute (KNMI) available online (http://climexp.knmi.nl/). Observation monthly periods are 36 Years January 1981 until December 2016.

The definition of these variables are: Niño12 is sea surface temperature (SST) anomalies in the the region of coastal South America (0-10S, 90W-80W), Nino34 is SST anomalies in the equatorial pacific 5°S to 5°N latitude and 170°W to 120°W longitude, Nino3 is SST anomalies in the pacific over the region 5°S to 5°N latitude 150°W to 90°W longitude, Nino4 is SST anomalies in the pacific(150°W-160°E and 5°N-5°S), DMI is differences in surface air pressure Tahiti and Darwin.

B. Data Preprocessing

Because we evaluate the relationship between features with rainfall in the monthly period, we must be aggregate local daily data to monthly data. For daily rainfall, we added the daily data per month. For minimum temperature, we chose a minimum value per month and for maximum temperature and maximum wind-speed, we chose a maximum value. Mean temperature wind-speed and solar radiation, we counted averages the daily data per month. Fig. 1 explains the characteristics of monthly rainfall data.

At the dataset, we have 4.27% missing value and we must do interpolation by replacing the empty data using the mean of the variable value for the month. The dataset is divided two subset namely training data set and testing data set. The first subset is consists of 396 records and the second is consisted of 36 records.

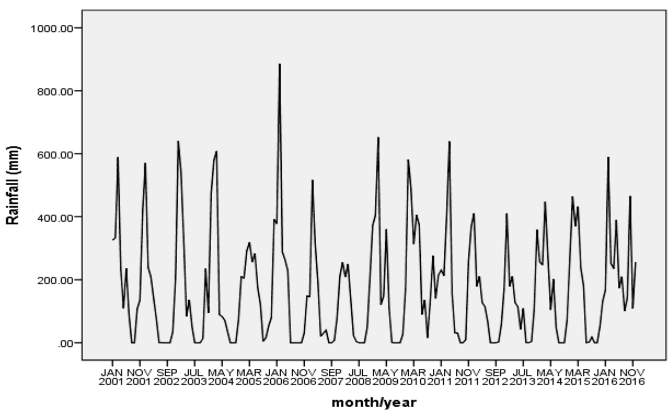

Fig. 1. Monthly Rainfall Time Series at Surabaya

III. Forecasting Methods

A. ARIMA

One of the time series forecasting which is popular and mostly used is ARIMA model. Based on Babu[12], AR model shows that there is a relation between a value in the present (Z_t) and values in the past (Z_{t-k}) added by random value e_t. MA model shows that there is a relation between a value in the present (Z_t) and residuals in the past (a_{t-k} with k=12, …). ARIMA(p, d, q) model is a mixture of AR(p) and MA(q) with a non-stationery data pattern and d differencing order. The form of ARIMA(p, d, q) is equation 1.

$$\phi_p(1-B)^d Z_t = \theta_q(B)a_t \quad (1)$$

Where p is AR model order q is MA model order d is difference order and

$$\phi_p(B) = (1-\phi_1 B - \phi_2 B^2 \dots - \phi_p B^p)$$
$$\theta_q(B) = (1-\theta_1 B - \theta_2 B^2 \dots - \theta_q B^q)$$

Generalization of ARIMA model for a seasonal pattern data which is written as ARIMA(p d q)(P D Q)^s is equation 2.

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)a_t \quad (2)$$

where s is seasonal period

$$\Phi_P(B^s) = (1-\Phi_1 B^s - \phi_2 B^{2s} \dots - \phi_P B^{Ps})$$
and $\Theta_Q(B) = (1-\Theta_1 B^s - \Theta_2 B^{2s} \dots - \Theta_Q B^{Qs})$.

Flow-chart for ARIMA model is describes at Fig. 2.

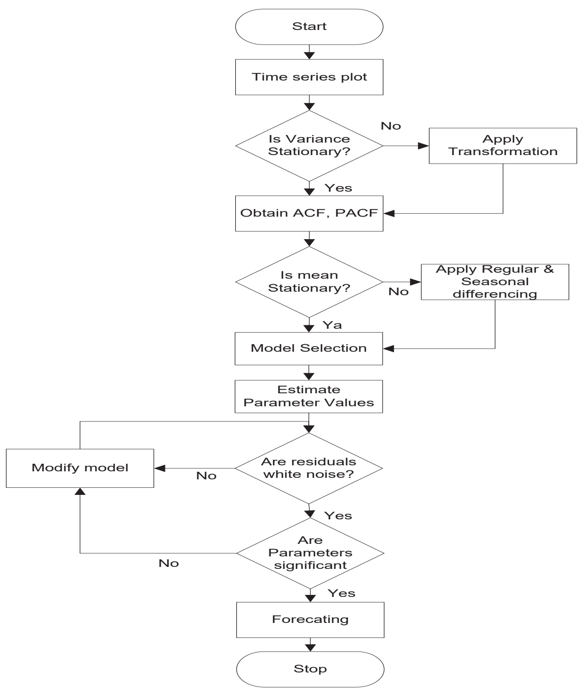

Fig. 2. Flow-chart of ARIMA modeling

B. Single Input Multi Transfer

Transfer function model is different from ARIMA model. ARIMA model is an univariate time series model, but transfer function is multivariate time series model. This means that ARIMA model relates the series only to its past. Besides the past series, transfer function model also relates the series to other time series as described fig. 3.

Fig. 3. Single-Input Multi Transfer

Transfer function models can be used to model single-output and multiple-output systems [5]. In the case of single-output model, only one equation is required to describe the system. It is referred to as a single-input transfer function model (equation 3).

$$Z_t = C + \frac{\omega_s(B)}{\delta_r(B)} B^b X_t + N_t \quad (3)$$

where $N_t = \frac{\theta(B)}{\phi(B)} a_t$

$$\omega_s(B) = \omega_1 B + \omega_2 B^2 \dots + \omega_s B^s$$

$$\delta_r(B) = \delta_1 B + \delta_2 B^2 \dots + \delta_r B^r$$

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 \dots - \phi_p B^p)$$

s is numerator, r is denominator and b is delay.

C. Multi-Input Transfer Function

A transfer function model may contain more than one input variable as in multiple regression models. In this paper, both single-input and multi-input transfer function is applied assuming that input and output series are both stationary the general form of a single-input transfer function model is equation 4. We can apply differencing technique if those variables are non stationary.

$$Z_t = C + \frac{\omega_{s1}(B)}{\delta_{r1}(B)} B^{b1} X_{1t} + \frac{\omega_{s2}(B)}{\delta_{r2}(B)} B^{b2} X_{2t} + \dots + \frac{\omega_{sm}(B)}{\delta_{rm}(B)} B^{b1} X_{mt} + N_t \quad (4)$$

D. Performance Model

The best model are evaluated using four performance measures, i.e. Root Mean Square Error (RSME), Mean Absolute Error (MAE), R square (R^2) and Correlation (r). A good model is to have small RMSE and MAE and have a large R^2 and r. The four performance measures are:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(Z_t - \hat{Z}_t)^2}{n}} \quad (5)$$

$$\text{MAE} = \frac{\sum_{t=1}^{n}|Z_t - \hat{Z}_t|}{n} \quad (6)$$

$$R_{sq} = 1 - \frac{\sum_{t=1}^{n}(Z_t - \hat{Z}_t)^2}{\sum_{t=1}^{n}(Z_t - \bar{Z}_t)^2} \quad (7)$$

$$\text{Correlation} = r = \sqrt{R^2} \quad (8)$$

IV. RESULT AND DISCUSSI

A. ARIMA model

Based on the flow-chart for ARIMA model (fig. 2), firstly, the time-series data is identified whether it has been stationary using the time-series plot. That monthly rainfall data has been stationary according to fig. 1. No transformation needed nor difference. Secondly, we identify time series stationarity ACF and PACF pattern of monthly rainfall. ACF correlogram in fig. 4 shows having a sine pattern which is suspected to be an AR seasonal model. This is reinforced by the PACF correlogram which shows sign1ificant correlation value at lag 1^{st} the 12^{th} and 24^{th}. The Correlogram (fig. 4 and 5) shows that the rainfall data is stationary because PACF has been cut off after the 1st lag. Based on the ACF and the PACF a number of ARIMA models are proposed (p d q)(P D Q)^S using training data.

Table I shows the six significant models in which the best model is the 3rd model as equation:

$$Y_t = -974.775 + 1.151Y_{t-12} - 0.152Y_{t-24} + 14.871X_t - 0.226e_{t-1} + 0.981e_{t-12} + e_t$$

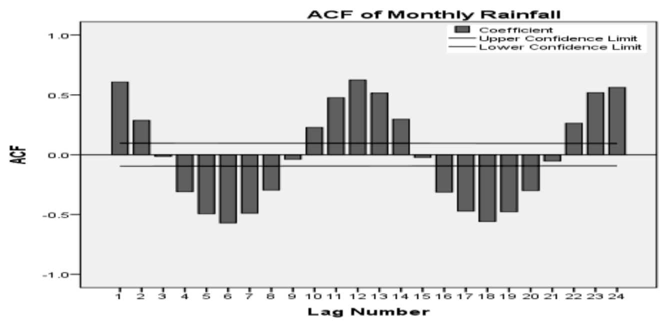

Fig. 4. ACF Correlogram of monthly rainfall

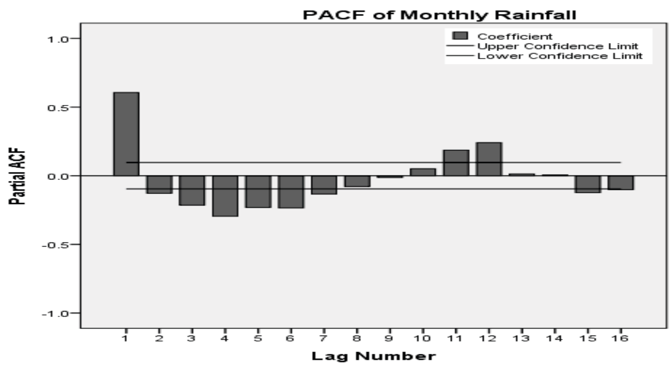

Fig. 5. PACF Correlogram of monthly rainfall

Residual from third model is diagnostic checking indicating that the model error has white noise according to fig. 6.

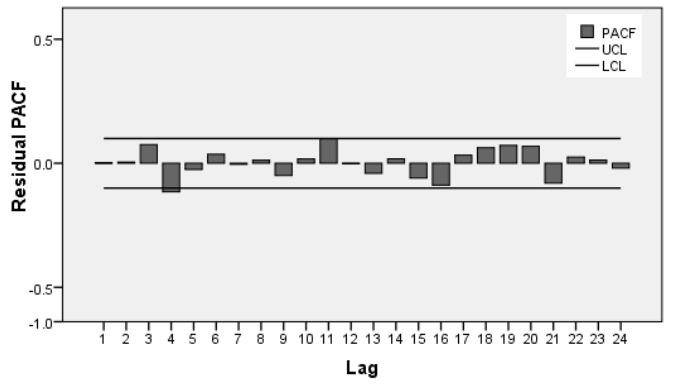

Fig. 6. PACF Corellogram of error model 3.

Since third model is the best, this model is used to predict the monthly rainfall value for the period 36 ahead (Jan 2014-Dec 2016). Based on data testing the performance of model 3 is est compared with other models. Performance testing for 6 models is summarized in table 2. While fig. 7 is the time series plot for real data, fit value (until /Dec 2013) and forecast data (Jan 2014- Dec 2016) using model 3.

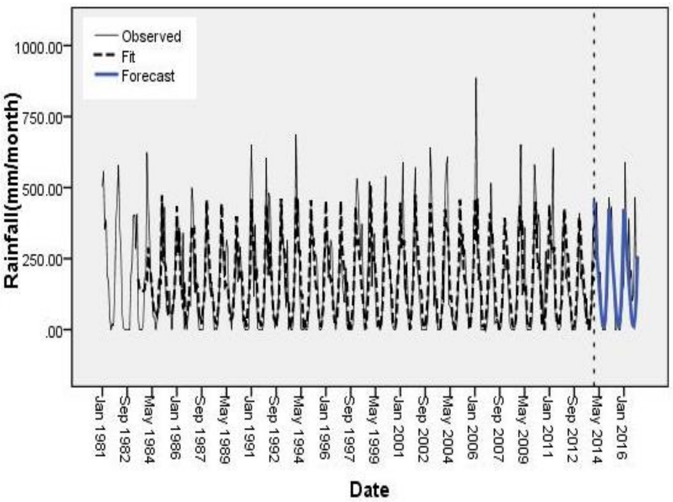

Fig. 7. Forecasting plot based model 3.

TABLE I. PERFORMANCE OF ARIMA BASED ON TRAINING DATA

No	Model	RMSE	MAE	R^2	R
1	$(1\ 0\ 0)(1\ 0\ 0)^{12}$	120.97	91.39	0.510	0.714
2	$(1\ 0\ 0)(2\ 0\ 0)^{12}$	117.92	88.56	0.535	0.731
3	$(1\ 0\ 0)(1\ 0\ 1)^{12}$	**106.34**	**75.79**	**0.622**	**0.789**
4	$(0\ 0\ 1)(2\ 0\ 0)^{12}$	118.51	90.00	0.530	0.728
5	$(0\ 0\ 1)(1\ 0\ 1)^{12}$	106.57	76.17	0.620	0.787
6	$(0\ 0\ 1)(2\ 0\ 1)^{12}$	106.58	76.02	0.621	0.788

Table II shows the performance testing of six models and shows that third model is best.

B. *Single-Input Transfer Function*

The rainfall forecasting model is improved by adding a single predictor variable. This model is called single input transfer function ARIMA. There are seven local predictor variables and five global predictor variables used as inputs in turn. Using 396 training data, it was found that the partial input predictors were humidity, max/mean temperature, wind-speed, solar radiation.

TABLE II. PERFORMANCE OF ARIMA FORECAST BASED ON TESTING DATA

No.	Model	RMSE	MAE	R^2	r
1	$(1\ 0\ 0)(1\ 0\ 0)^{12}$	140.05	111.1	0.26	0.513
2	$(1\ 0\ 0)(2\ 0\ 0)^{12}$	138.46	106.6	0.29	0.543
3	$(1\ 0\ 0)(1\ 0\ 1)^{12}$	122.80	83.45	0.45	0.670
4	$(0\ 0\ 1)(2\ 0\ 0)^{12}$	140.56	107.2	0.28	0.525
5	$(0\ 0\ 1)(1\ 0\ 1)^{12}$	122.67	83.30	0.45	0.671
6	$(0\ 0\ 1)(2\ 0\ 1)^{12}$	**122.66**	**83.11**	**0.45**	**0.671**

Two other local variables and five global predictor variables are not significant input predictors. The best model based on training data for single input transfer function (SITF) is CH → Rainfall (Table III) with r=0.84. While based on data testing, the best RMSE and MAE are model 5 (SR → Rainfall), but RH →Rainfall has the best R^2 and r (Table IV).

TABLE III. PERFORMANCE OF SI-TF MODEL BASED ON TRAINING DATA

Input –ARIMA	(b s r)	$(b\ s\ r)^{12}$	RMSE	MAE	R^2
RH $(1\ 0\ 0)(0\ 1\ 1)^{12}$	(0 0 0)	(0 0 0)	**92.99**	**66.34**	**0.71**
MaxT$(1\ 0\ 0)(1\ 0\ 1)^{12}$	(1 0 0)	(0 0 0)	105.69	74.92	0.62
MeanT$(1\ 0\ 0)(1\ 0\ 1)^{12}$	(1 0 1)	(0 0 0)	105.1	74.55	0.62
WS $(1\ 0\ 1)(1\ 0\ 1)^{12}$	(1 0 2)	(0 0 1)	105.49	75.04	0.62
SR $(2\ 0\ 0)(0\ 1\ 1)^{12}$	(0 0 0)	(0 0 0)	103.28	72.89	0.64

TABLE IV. PERFORMANCE SI-TF MODEL BASED ON TESTING DATA

No.	Input	RMSE	MAE	R^2	r
1	RH	156.57	132.80	**0.608**	**0.780**
2	Max T	163.54	140.26	0.461	0.679
3	Mean T	138.67	120.13	0.531	0.729
4	WS	163.77	139.27	0.367	0.606
5	SR	**119.1**	**82.54**	0.501	0.708

C. Multi-Input Transfer Function

Single input model can be developed into multi input transfer function by using five predictor variables simultaneously, (humidity, max/mean temperature, wind-speed, solar radiation) → rainfal. The result of this multi input modeling is that wind-speed as a predictor local variable is not significant. By using four predictor variables simultaneously, the performance model based on training data is better than the previous two models. Table V summarizes the performance of the multi input transfer function model for training data, with r = 0.85.

TABLE V. PERFORMANCE OF MI-TF MODEL BASED ON TRAINING DATA

Rainfall	Input	(b s r)	$(b\ s\ r)^{12}$	RMSE	MAE	R^2
$(1\ 0\ 0)(1\ 0\ 1)^{12}$ or $(0\ 0\ 1)(2\ 0\ 1)^{12}$	RH MaxT MeanT SR	(0 0 0)	(0 0 0)	91.49	65.02	0.72
		(1 0 0)	(0 0 0)			
		(1 0 1)	(0 0 0)			
		(2 0 0)	(0 0,0)			

Table VI summarizes the performance of the multi input transfer function model. From this result it can be concluded that multi input input transfer function gives best result, applicable for training and testing data.

TABLE VI. PERFORMANCE OF 3 MODELS ON TRAINING &TESTING DATA

Model	RMSE	MAE	R^2	r
Training				
ARIMA	106.34	75.79	0.622	0.79
SI-TF:	92.99	66.34	0.71	0.84
MI-TF:	**91.49**	**65.02**	**0.72**	**0.85**
Testing:				
ARIMA	122.66	83.11	0.45	0.67
SI-TF	119.10	82.54	0.61	0.78
MI-TF	**100.29**	**73.46**	**0.616**	**0.785**

V. CONCLUSIONS

From comparison among three methods of rainfall forecasting, it is found that seasonal influence is very significant to the model with 12 month-period, both ARIMA and transfer function. This means the three models are able to detect any seasonal influence. The ARIMA univariate model, although it produces the appropriate model, has the lowest performance compared to single input and multi input transfer function. This suggests that rainfall forecasting models require predictors of exogenous variables.

The best exogenous variables are humidity, solar radiation, maximum temperature and mean temperature simultaneously where the model accuracy reaches 72%, and the correlation between real data and forecast data is 0.85. Global predictor variables do not show a significant effect on model performance.

REFERENCES

[1] E. Aldrian, "Simulations of Indonesian Rainfall with a Hierarchy of Climate Models," no. 92, 2003.

[2] J. Sánchez-Monedero, S. Salcedo-Sanz, P. A. Gutiérrez, C. Casanova-Mateo, and C. Hervás-Martínez, "Simultaneous modelling of rainfall occurrence and amount using a hierarchical nominal-ordinal support vector classifier," *Eng. Appl. Artif. Intell.*, vol. 34, pp. 199–207, 2014.

[3] J. Abbot and J. Marohasy, "Input selection and optimisation for monthly rainfall forecasting in queensland, australia, using artificial neural networks," *Atmos. Res.*, vol. 138, pp. 166–178, 2014.

[4] M. Han, W. Ren, and X. Liu, "Joint mutual information-based input variable selection for multivariate time series modeling," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 250–257, 2015.

[5] B. Widjanarko Otok, "Development of Rainfall Forecasting Model in Indonesia by using ASTAR, Transfer Function, and ARIMA Methods," *Eur. J. Sci. Res.*, vol. 38, no. 3, pp. 1450–216, 2009.

[6] J. Farajzadeh, A. Fakheri Fard, and S. Lotfi, "Modeling of monthly rainfall and runoff of Urmia lake basin using 'feed-forward neural network' and 'time series analysis' model," *Water Resour. Ind.*, vol. 7–8, no. 2014, pp. 38–48, 2014.

[7] A. Nugroho and B. H. Simanjuntak, "ARMA (Autoregressive Moving Average) Model for Prediction of Rainfall in Regency of Semarang - Central Java - Republic of Indonesia," *Int. J. Comput. Sci. Issues*, vol. 11, no. 3, pp. 27–32, 2014.

[8] A. Nugroho, S. Hartati, and K. Mustofa, "Vector Autoregression (Var) Model for Rainfall Forecast and Isohyet Mapping in Semarang–Central Java–Indonesia," *... J. Adv. Comput. Sci. ...*, vol. 5, no. 11, pp. 44–49, 2014.

[9] M. Fuentes, B. Reich, and G. Lee, "Spatial-temporal mesoscale modeling of rainfall intensity using gage and radar data," *Ann. Appl. Stat.*, vol. 2, no. 4, pp. 1148–1169, 2008.

[10] J.-I. Hamada, M. D. Yamanaka, J. Matsumoto, S. Fukao, P. A. Winarso, and T. Sribimawati, "Spatial and Temporal Variations of the Rainy Season over Indonesia and their Link to ENSO.," *J. Meteorol. Soc. Japan*, vol. 80, no. 2, pp. 285–310, 2002.

[11] A. di Piazza, F. Lo Conti, L. V. Noto, F. Viola, and G. La Loggia, "Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 13, no. 3, pp. 396–408, 2011.

[12] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data," *Appl. Soft Comput.*, vol. 23, pp. 27–38, 2014.

This page intentionally left blank