TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 8, August 2014, pp. 6423 ~ 6430
DOI: 10.11591/telkomnika.v12i8.6195

Downscaling Modeling Using Support Vector Regression for Rainfall Prediction

Sanusi*[1], Agus Buono[2], Imas S Sitanggang[3], Akhmad Faqih[4]
[1,2,3]Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University, 16680 Bogor, Indonesia, Ph/Fax. +62-251-628448/622961
[4]Department of Geophysics and Meteorology, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University, 16680 Bogor, Indonesia, Ph/Fax. +62-251-628448/622961
Corresponding author, e-mail: sanusiumarhasan@gmail.com*[1], pudesha@yahoo.co.id[2],
imas.sitanggang@gmail.com[3], akhmadfa@ipb.ac.id[4]

Abstract

Statistical downscaling is an effort to link global scale to local scale variable. It uses GCM model which usually used as a prime instrument in learning system of various climate. The purpose of this study is as a SD model by using SVR in order to predict the rainfall in dry season; a case study at Indramayu. Through the model of SD, SVR is created with linear kernel and RBF kernel. The results showed that the GCM models can be used to predict rainfall in the dry season. The best SVR model is obtained at Cikedung rain station in a linear kernel function with correlation 0.744 and RMSE 23.937, while the minimum prediction result is gained at Cidempet rain station with correlation 0.401 and RMSE 36.964. This accuracy is still not high, the selection of parameter values for each kernel function need to be done with other optimization techniques.

Keywords: *statistical downscaling, general circulasi models, support vector regression, rainfall in dry season*

1. Introduction

In some recent years ago, many efforts have already done to explore the effect of climate variety whether in a big scale or climate change toward the variability of rainfall in the worldwide [1]. The climate variety especially rainfall in Indonesia mostly influenced by global phenomenon such as El-Nino and Southern Oscillation (ENSO), ENSO is conventionally identified as ocean temperature warming in eastern Pacific [2]. Indian Ocean Dipole (IOD), IOD as a modus of tropical physic in Indian Ocean is strongly believed as a main effect which causes dryness in Indonesia [3]. Madden Julian Oscillation (MJO), MJO as a global phenomenon influences the climate in western of Indonesia [4]. This phenomenon also happens in Indramayu. It is one of Indonesia district which has monsoon rain and as a central production of agriculture particularly rice [5]. The main factors cause crop failures in Indramayu are dryness (79.8%), pest attack (15.6%) and float (5.6%) [6].

One of instruments which can be used to observe the indication of climate variability is General Circulation Mode [7]. It can be known that GCM has an intense relationship between big scale climate and whether on local scale for rainfall prediction [8], [9]. Simulated rainfall pattern from the various models of GCM is able to give basic information that needed to the future development [10]. However, GCM data is considered to the low of resolution and global scale which difficult to be used in doing prediction because local climate needs high resolution, but GCM is still can be used if it mixed to the downscaling technique.

Many models that already used to predict climate in GCM and SD such as Buono *et al* (2010) [11] statistical downscaling modeling using Artificial Neural Networks (ANN) for prediction monthly rainfall in Indramayu. In addition, Wigena (2006) [12] statistical downscaling model with Regression Projection Persuit (PPR) to forecast the rainfall (monthly rainfall case in Indramayu). This study uses Support Vector Regression on downscaling model to predict the rainfall in dry season.

Received April 1, 2014; Revised June 3, 2014; Accepted June 15, 2014

Statistical downscaling is defined as transfer function that describes functional relationship of global atmospheric circulation with local climate elements [13]. Figure 1 is process illustration of downscaling statistical.

$$Y_{t,p} = f(X_{t,q,s,g}) \quad (1)$$

Where,

Y = local climate variable
X = GCM output variable
t = time period
p = many of Y variable
q = many of X variable
s = many of atmosphere layer
g = GCM domain

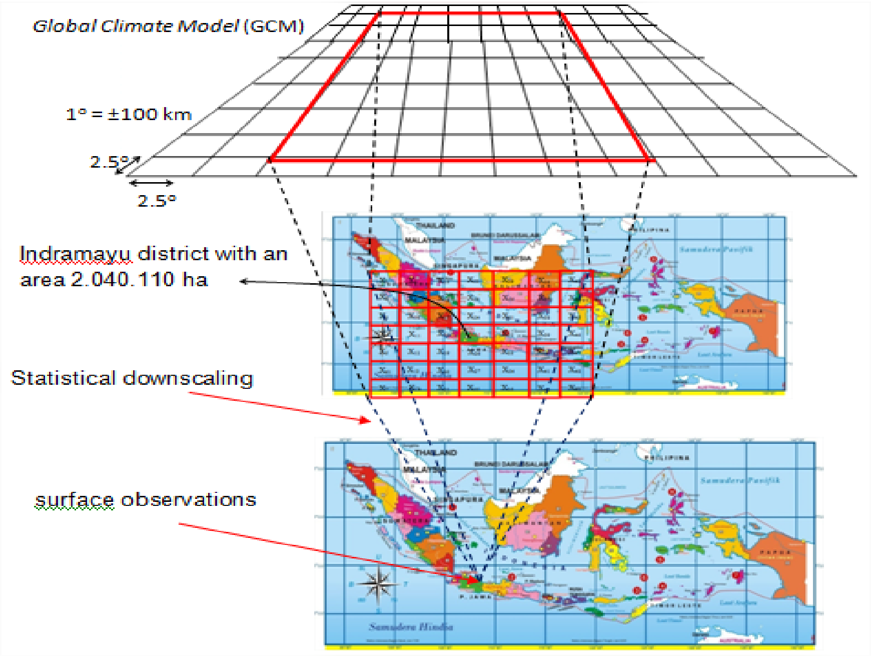

Global Climate Model (GCM)
1° = ±100 km
2.5°
2.5°
Indramayu district with an area 2.040.110 ha
Statistical downscaling
surface observations

Figure 1. Statistical Downscaling Illustration

1.2. Support Vector Regression

Support Vector Regression (SVR) is the expansion of Support Vector Machine (SVM). SVM used to solve clarification problem, while SVR used to regression case. SVR is a method that can overcome overfitting, so that it will result better performance [14].

Suppose we have a set of data as much as ℓ set training data in a formula:{$\chi = xi,yi$ with $i=1,\ldots,\ell$, by x input data = {x_1, x_2, x_3, ...,n} $\subseteq \Re N$ and the corresponding output as $\{y = [y_i, \ldots, y_\ell] \subseteq \Re\}$. When ε value is equal as 0, we will get a perfect regression. Suppose we have a function as regression line below:

$$f(x) = w \cdot \phi(x) + b \quad (2)$$

By $\phi(x)$ shows a point in feature space F the mapping result of x in input space. Coefficient of w and b are estimated by minimizing the risk function that describes in the following formulation:

$$\min \frac{1}{2} \parallel w \parallel^2 + C \frac{1}{\ell} \sum_{i=1}^{\ell} L_{\in}(y_i, f(x_i)) \quad (3)$$

Depends on

$$y_i - w\phi(x_i) - b \leq \varepsilon$$
$$w\phi(x_i) + b - y_i \leq \varepsilon, i = 1, 2, 3, \dots, \ell$$

With,

$$L_\varepsilon(y_i, f(x_i)) = \begin{Bmatrix} |y_i - f(x_i)| - \varepsilon, |y_i - f(x_i)| \geq \varepsilon \\ 0 \qquad\qquad\qquad , \text{to the others} \end{Bmatrix}$$

By minimizing $\| w \|^2$ will make the function as thin as possible, as a result the capacity function can be controlled. ε-insensitive loss function required to minimize norm from w achieve better generalization to regression function f(x). That is why we have to solve the following problem:

$$\min \frac{1}{2} \| w \|^2 \tag{4}$$

Depends on:

$$y_i - w\phi(x_i) - b \leq \varepsilon$$
$$w\phi(x_i) + b - y_i \leq \varepsilon, i = 1, 2, 3, \dots, \ell$$

Assume the function of f(x) which can approximate to all of these points (x_i, y_i). Then, we will get a cylinder as describe in Figure 2.

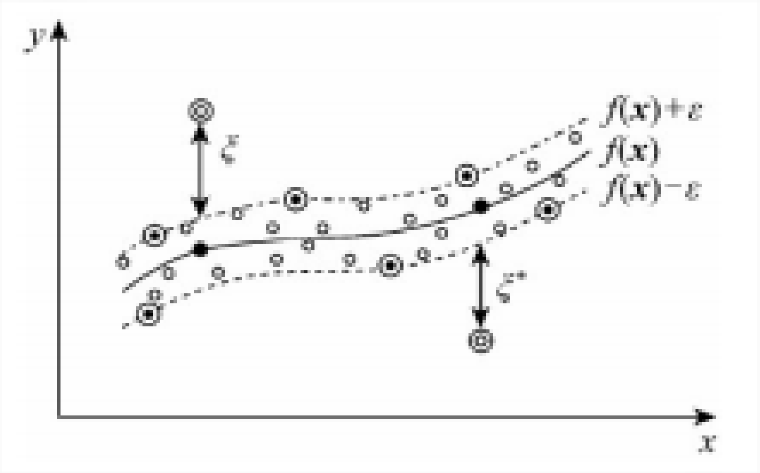


Figure 2. Regression Function at SVR [15]

Accuracy of ε in this case we assume that all points in the range f ± ε (feasible). In the case of ineligibility, where there are some points that may be out of range f ± ε, we need to add variable of slack ξ, ξ^*. Furthermore, the optimization problem can use the following formula:

$$\min = \frac{1}{2} \|w\|^2 + C \sum_{\ell=1}^{\ell} (\xi_i + \xi_i^*) \tag{5}$$

Depends on:

$$y_i - w^T\phi(x_i) - \xi_i - b \leq \varepsilon, i = 1, 2, 3, \dots, \ell$$
$$w\phi(x_i) - y_i - \xi_i^* + b \leq \varepsilon, i = 1, 2, 3, \dots, \ell$$
$$\xi, \xi^* \geq 0$$

The constant of C > 0 determined the bargaining between the thinness of function f and the upper limit of deviation that more than ε was still tolerated. ε was comparable to the accuracy of the approximation of the training data. The highest value of ε was related to ξ_i^* that has small and low approximation accuracy. The highest value for variable ξ_i^* will make empirical errors which have a considerable influence on the regularization factor. In SVR support vector there was the training data which located out of f from the decision function.

By C was determined by user, $K(x_i, x_j)$ was dot-product kernel that identified as $K(x_i, x_j) = \phi^T(x_i)\,\phi^T(x_j)$, by using Lagrange multipliers and optimalization condition, The regression function was formulated explicitely in the following formula:

$$f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\,K(x, x_i) + b \tag{6}$$

Before doing training and test of SVR, it is better for us to decide parameter value of C, ε to the function of Linear Kernel and C parameter, ε, and γ to RBF kernel function.

2. Research Method

This study was undertaken in several phases. All of those phases can be seen in the following figure Figure 3.

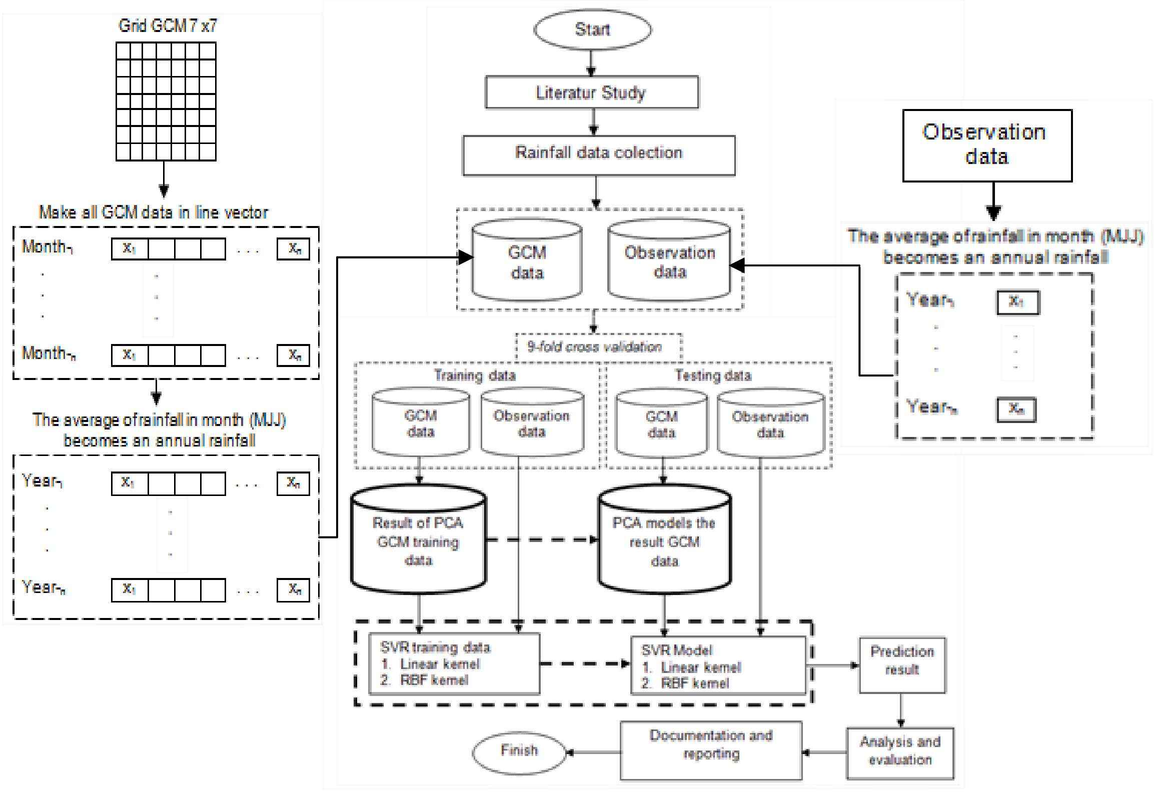


Figure 3. Research Flowchart

The beginning of this study was literature review; it used in order to understand all problems that will be researched. The data used in this research is secondary data divided to GCM hindcast data result (used as clarify variable) and data of rainfall observation (used as respond variable). Result of GCM hindcast data was acquired from the Climate Information Tool Kit (CLIK) APEC Climate Center (APCC) as the rainfall data and type of ASCII file which consists of 6 models with a resolution grid of latitude and longitude 2.5^{0}x2.5^0, data accessed from the website CLIK APCC (http://clik.apcc21.org), as well as two models of GCM hindcast rainfall obtained from the website of the International Research Institute Data Library (IRIDL) (http://iridl.ldeo.columbia.edu), as data of Climate Prediction Center(CPC) Unified Gauge-Based Analysis of Global Daily Precipitation from The International Research Institute for Climate and Society (IRI) and TSV file type with a grid resolution of latitude and longitude 0.5^{0}x0.5^{0}. Hindcast GCM data used to build prediction model in 3 different months: May, June, and July (MJJ) from the year of 1982-2008 (27 years) every model at every rainfall station. In this study, there are 8 GCM hindcast rainfalls to build prediction model as shown in Table 1.

The data of rainfall observation (respond variable) is the average value of seasonal rainfall at every rainfall station in Indramayu by longitudinal position of$107^{o}52^{'}$-$108^{o}36^{'}$BT and $6^{o}15^{'}$-$6^{o}40^{'}$LS, it was obtained from the measurement and test that performed by Meteorology Department in Indramayu. There were 15 observation stations used as shown in Table 2. The data of rainfall observation was used 3 months: May, June, July (MJJ) from the year of 1982-2008 (27 years) at every rainfall station.

Data of GCM was cropped in grid of 7x7 and then make all of GCM data model to the line vector; Next, average rainfall of data GCM and observations to be the annual rainfall. Furthermore, distribute training and test data by using 9-fold cross Validation, 9 is divided due to the number of year and redone in nine times. The data PCA is necessary to be done because it can avoid the double linear data in GCM model and to save computing time during training and testing the SVR model. Reduction process is held by taking one or more major components with diversity of ≥98%. Finally the SVR training and testing can be done.

Tabel 1. The Data of GCM Hindcest Rainfall and its Founders

No	Model Name	Ensemble	Institution	Sources	References
1	GCPS T63T21	4	Korea	http://clik.apcc21.org	[16]
2	GDAPS T106L21	20	Korea	http://clik.apcc21.org	[16]
3	CMC1-CanCM3	120	Columbia	http://iridl.ldeo.columbia.edu	[17], [19]
4	CanCM3-AGCM3	10	Canada	http://clik.apcc21.org	[16]
5	GFDL-CM2P1	120	Columbia	http://iridl.ldeo.columbia.edu	[17], [19]
6	NASA-GSFC L34	8	U.S.A	http://clik.apcc21.org	[16]
7	METRI AGCM L17	10	Korea	http://clik.apcc21.org	[16]
8	PNU	5	Korea	http://clik.apcc21.org	[16]

Tabel 2. The Name and Location of the 15 Rainfall Observation Stations in Indramayu

Y	Station Name	LS	BT	Y	Station Name	LS	BT
Y_1	Bangkir	-6.336	108.325	Y_9	Ujungaris	-6.457	108.287
Y_2	Bulak	-6.338	108.116	Y_{10}	Loh berner	-6.406	108.282
Y_3	Cidempet	-6.354	108.246	Y_{11}	Sudimampir	-6.402	108.366
Y_4	Cikedung	-6.492	108.185	Y_{12}	Juntinyuat	-6.433	108.438
Y_5	Losarang	-6.398	108.146	Y_{13}	Krangkeng	-6.503	108.483
Y_6	Sukadana	-6.535	108.300	Y_{14}	Bondan	-6.606	108.299
Y_7	Sumurwatu	-6.337	108.325	Y_{15}	Kedokan Bunder	-6.509	108.424
Y_8	Tugu	-6.433	108.333				

3. Results and Analysis

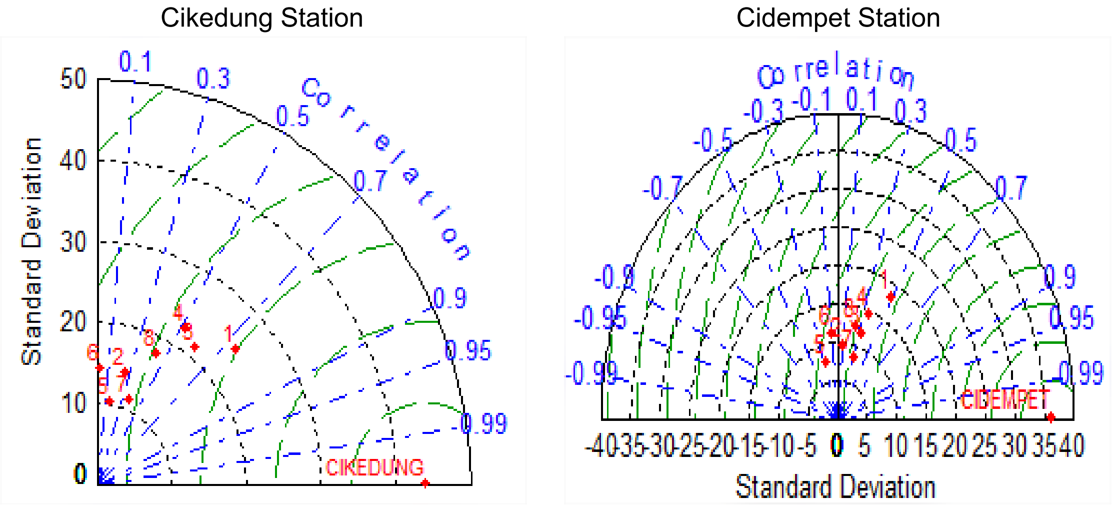
Downscaling model by using SVR to predict the rainfall in dry season with clarify variable in model of GCM and observation of rainfall as respond variable, All of those data were used at every 15 rainfall stations in Indramayu. Here are the results of the prediction of the model GCM rainfall averaged as shown in Table 3.

Based on the prediction result on Table 3, it can be said that the result will be better if it has a high correlation while RMSE in low value. On the kernel linear function the high correlation value was obtained at Cikedung rainfall station. On the other hand, the low correlation value was gotten at Cidampet rainfall station. Overall, it can be concluded that result production by using kernel linear function was better than RBF kernel function. It was marked by the correlation value or RMSE value in every rainfall station.

Tabel 3. The Average Correlation of the Prediction Result by using GCM Model Data and RMSE Values between Rainfall Observation in Indramayu

No	Station	Kernel Linear		Kernel RBF	
		Correlation	RMSE	Correlation	RMSE
1	Bangkir	0.578	62.269	0.562	67.799
2	Bulak	0.684	26.052	0.345	30.298
3	Cidempet	**0.401**	**36.964**	0.241	35.353
4	Cikedung	**0.744**	**23.937**	0.538	42.483
5	Losarang	0.721	26.955	0.556	32.823
6	Sukadana	0.419	30.517	0.528	31.287
7	Sumurwatu	0.670	36.918	-0.053	42.855
8	Tugu	0.651	28.449	0.472	32.258
9	Ujungaris	0.515	29.653	0.422	32.261
10	Lohbener	0.675	32.349	0.579	35.478
11	Sudimampir	0.514	55.424	0.472	57.634
12	Juntinyuat	0.611	44.384	0.648	49.783
13	Kedokan Bunder	0.726	39.267	0.696	43.202
14	Krangkeng	0.655	43.335	0.414	49.422
15	Bondan	0.681	24.730	0.208	27.580

The best GCM model was in Taylor chart that closer to the observation point. By looking at standard deviation, RMSE and correlation, observation point is the standard deviation of data point at a particular location [20]. There are 8 explanation of GCM models we can find at Taylor chart, they are: 1. CMC1-CanCM3, 2. GDAPS T106L21, 3. GFDL-CM2P1, 4. GCPS T63T21, 5. CanCM3-AGCM3, 6. METRI AGCM L17, 7. NASA-GSFC L34, 8. PNU. Here is Taylor chart for GCM model at Cikedung and Cidempet rainfall station as shown in Figure 5.


Cikedung Station

Cidempet Station

Figure 5. Taylor Chart for GCM Model

Based on the chart in Figure 5, it was known that Cikedung rainfall station was at standard deviation about ±44 and RMSE value ±30. The 1 model was potentiality to be the best model in this location if it compared to another model while Cidempet rainfall station was at ±36 standard deviation. The 1 model became the best model in this location if it compared to another model. But, the 1 model at Cidempet station was not as better as 1 model at Cikedung station, it was caused by the 1 model at Cidempet station has ±32 RMSE value. The overall of linear kernel function was better than RBF kernel function.

4. Conclusion

To sum it up, the models which were resulted to predict the rainfall in dry season will be better if it looked from the average of prediction result or the error average. The best correlation value was obtained at Cikedung rainfall station in 0.744 correlation value and 23.937 RMSE while the lowest linear kernel function was gained at Cidempet rainfall station in 0.401 correlation value and 36.964 RMSE. The kernel function of RBF was not included to the best function because the result prediction was lower than linear kernel function. It can be seen from the correlation value or RMSE on RBF kernel function.

Suggestion to the next research, downscaling model of GCM model data can be applied in order to predict the rainfall in dry season by using Support Vector Regression. The utilization of GCM grid can be used besides grid of 7x7. The accuracy was not high yet, and then the selection of parameter values for each kernel function needs to be performed with other optimization techniques.

References

[1] Karamouz M, Fallahi M, Nazif S, Farahani, RM. Long Lead Rainfall Prediction Using Statistical Downscaling and Artificial Neural Network Modeling. *Archive of SID. Sharif University of Technology*. 2009; 16(2): 165-172.

[2] Chen TS, Yang CT, Kuo MC, Kuo HC, Yu SP. Probabilistic Drought Forecasting in Southern Taiwan Using El Niño-Southern Oscillation Index. *Terr. Atmos. Ocean. Sci.* 2013; 24(5): 911-924.

[3] Ashok K, Guan Z, Yamagata T. *A Look at the Relationship between the ENSO and the Indian Ocean Dipole*. J Meteorological Society. 2003; 18(1): 41-56

[4] Evana L, Effendy S, Hermawan E. Pengembangan Model Prediksi *Madden Julian Oscillation* (MJO) Berbasis pada Hasil Analisis Data *Real Time Multivariate* MJO (RMM1 dan RMM2). *J. Agromet.* 2008; 22 (2): 144-159.

[5] Zein. Pemodelan Backpropagation Neural Networks dan Probabilistic Neural Network untuk Pendugaan Awal Musim Hujan Berdasarkan Indeks Iklim Global. Thesis. Bogor: Postgraduate Bogor Agriculture University; 2006.

[6] Estiningtyas W. Pengembangan Model Asuransi Indeks Iklim untuk Meningkatkan Ketahanan Petani Padi dalam Menghadapi Perubahan Iklim. Dissertation. Bogor: Postgraduate Bogor Agriculture University; 2012

[7] Villages RJ, Jarvis A. Downscaling Global Circulation Model Outputs: The Delta Method Decision and Policy Analysis Working. *J Centro International de Agricultura Tropical International Center for Tropical Agriculture*. 2010; 1: 1-18.

[8] Liu Y, Fan K. A New Statistical Downscaling Model for Autumn Precipitation in China. *J. Climatol*. 2012; DOI: 10.1002/joc.3514.

[9] Kannan S, Ghosh S. Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. Springer, Department of Civil Engineering, Indian Institute of Technology Bombay. India, Spinger. 2010; DOI 10.1007/s00477-010-0415-y.

[10] Faqih A. Rainfall Variability in the Austral-Indonesian Region and the Role of Indo-Pacific Climate Drivers. Dissertation University of Southern Queensland. 2010.

[11] Buono A, Faqih A, Boer R, Santikayasa IP, Ramadhan A, Muttaqien MR, Asyhar A. A Neural Network Architecture for Statistical Downscaling Technique: A Case Study in Indramayu District. Publication in International Conference, The Quality Information for Competitive Agricultural *Based Production System and Commerce (AFITA)*. 2010.

[12] Wigena HA. Pemodelan statistical downscaling dengan regression projection persuit untuk peramalan curah hujan (kasus curah hujan bulanan di Indramayu). Dissertation. Bogor: Postgraduate Bogor Agriculture University. 2006.

[13] Sutikno. *Statistical Downscaling Luaran GCM dan Pemanfaatannya untuk Peramalan Produksi Padi.* Dissertation. Bogor: Postgraduate Bogor Agriculture University. 2008.

[14] Smola A, Schölkopf B. A Tutorial on Support Vector Regression. NeuroCOLT, *Technical Report NC-TR-98-030*, Royal Holloway College, University of London, UK. 2004; 199–222.

[15] Arampongsanuwat S, Meesad P. Prediction of PM_{10} using Support Vector Regression. International Conference on Information and Electronics Engineering, IACSIT Press. Singapore. 2011; 6.

[16] An KH, Heo Jin Y, Hameed SN. *CLIK 2.0 CLimate Information tool Kit User Manual*. APEC Climate Center. 2010.

[17] Xie P, Yatagai A, Chen M, Hayasaka T, Fukushima Y, Liu C, Yang S. A gauge-based analysis of daily precipitation over East Asia. *Journal of Hydrometeorology.* 2008; (8): 607-627.

[18] Chen MW, Shi P, Xie VBS, Silva VE, Kousky R, Higgins W, Janowiak JE. Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys*. 2008. Res. 113, D04110, doi:10.1029/2007JD009132.

[19] Chen M, Xie P. *CPC Unified Gauge-based Analysis of Global Daily Precipitation*. Western Pacific Geophysics Meeting, Cairns, Australia. 2008.

[20] Taylor KE. Summarizing multiple aspect of model performance in a single diagram. J Geophysical Research: Atmospheres. 2001; 106(D7): 7183-7192.