Support Vector Regression Modelling for Rainfall Prediction in Dry Season Based on Southern Oscillation Index and NINO3.4

Gita Adhani[1], Agus Buono[1], Akhmad Faqih[2]
[1]Department of Computer Science, [2] Department of Geophysics and Meteorology
Faculty of Mathematics and Natural Sciences, Bogor Agricultural University
Email : adhani.gita@gmail.com, pudesha@gmail.com, akhmadfaqih@gmail.com

***Abstract*—Various climate disasters in Indonesia are mostly related to the El Nino Southern Oscillation (ENSO) phenomenon. The variability of climate especially rainfall is strongly related to this phenomenon. Southern Oscillation Index (SOI) and sea surface temperature anomaly (SSTA) at Nino3.4 region are two common indicators used to monitor phenomenon of El Nino and La Nina. Furthermore, SOI and NINO SSTA can be the indicator to find the rainfall probability in a particular season, related to the existing condition of climate irregularities. This research was conducted to estimate the rainfall during dry season at Indramayu district. The basic method used in this study was Support Vector Regression (SVR). Predictors used were SOI and NINO3.4 sea surface temperature (SST) data. The experiments were conducted by comparing the model performance and prediction results. The training set was clustered in advance and then SVR model was generated using RBF kernel based on their clustering result. This research obtained an SVR model with correlation coefficient of 0.76 and NRMSE error value of 1.73.**

I. INTRODUCTION

Climate is one of natural ecosystem components that has a major influence on the various sectors of human life. Indonesia as an agricultural country is dependent on climate condition and weather. Climate and weather are crucial factors to successes agricultural and plantation. Knowledge of climate patterns and weather can help in making decisions cropping patterns and plant varieties appropriate in different areas.

Various climate disasters in Indonesia are mostly related to the El Nino Southern Oscillation (ENSO) phenomenon. Climate variability, especially rainfall, is strongly related to this phenomenon. Generally, El Nino impact on rainfall decreasing or even drought, otherwise La Nina influences on rainfall increasing which can cause flooding [1].

La Nina causes cummulation of air mass that contains a lot of water vapor in the Indonesia atmosphere thus potency of rain clouds forming enhances. As a result, although the middle of 2010 dry season, it still could be raining in many regions with low up to high intensity [2].

El Nino phenomenon gives more serious impact than La Nina. El Nino causes rainfall in most area in Indonesia reduced. This rainfall decreasing rate is really dependent on intensity and El Nino duration. El Nino is noted once caused long-term drought in Indonesia. Rainfall information during dry season is greatly needed in agricultural and plantations. Rainfall forecasting during dry season can be used as information for farmers to mitigate any cases that can be happened like preproduction drought that lead to crop failure.

This research porpose to forecast rainfall during dry season by took case study in Indramayu region using Support Vector Regression (SVR) and related variables used are Southern Oscillation Index (SOI) and sea surface temperature (SST) at NINO 3.4 region. SVR is Support Vector Machine (SVM) is used for regression case.

Regression is one of common season prediction methods. Support Vector Machine (SVM) is used to solve barriers in statistical regression analysis. Linear regression based on several assumptions thus there can not always suitable to the existing data set characteristics. Formers research that applied SVR method by Larasati (2012) about rainy season onset prediction used Southern Oscillation Index (SOI) data [3]. Agmalaro (2011) also studied about statistical

Manuscript received June 28, 2013. This work was supported in part by the Computer Science Department of Bogor Agricultural University (BAU), Center for Climate Risk and Opportunity Management in Southeast Asia and Pacific (CCROM-SEAP), Bogor Agricultural University (BAU), United States Agency for International Development (USAID).

Gita Adhani is with the Computer Science Department, Bogor Agricultural University, PO BOX 16680 INDONESIA (corresponding author to provide phone: +6285366015984; e-mail: adhani.gita@gmail.com).

Agus Buono is with the Computer Science Department, Bogor Agricultural University, PO BOX 16680 INDONESIA (e-mail: pudesha@gmail.com).

Akhmad Faqih is with the Geophysics and Meteorology Department, Bogor Agricultural University, PO BOX 16680 INDONESIA (e-mail: akhmadfaqih@gmail.com).

downscaling modeling of GCM data using support vector regression to predict monthly rainfall in district of Indramayu. The results is quite good to predict rainfall in normal conditions, however it is neither extreme case [4].

II. METHOD

A. Problem Identification and Formulation

SOI and NINO sea surface temperature anomaly (SSTA) are used as indication in monitoring of El Nino and La Nina phenomenon that is commonly called by El Nino Southern Oscillation (ENSO). Southern Oscillation Index (SOI) is anomalies of air pressure difference in the Tahiti surface in Polynesia islands, French, with Darwin surface, Australia. The sustainable of SOI negative values below -8 shows El Nino phenomenon while SOI positive values above 8 shows La Nina phenomenon [5]. The more negative SOI values mean stronger warm event, whereas positive SOI values the stronger event cold event [6].

NINO is an index of sea surface temperature. NINO is obtained by taking the average value of the surface temperature in a given area. There are 4 NINO areas according to IRI [7], namely NINO1+2, NINO3, NINO3.4, and NINO4. NINO1+2 region is located between 0 ° - 10 ° S and 80 ° - 90 ° W. This area was first rise in temperature when El Nino occurs. NINO3 region lies in the middle of the Pacific Ocean between 5 ° N - 5 ° S and 90 ° - 150 ° W which is the zone most closely related to El Nino conditions. NINO3.4 region located between the equator 5 ° S - 5 ° N and 170 ° - 120 ° W and have a great variability in the time scale El Nino. NINO4 lies in the western of Pacific Ocean between 5 ° N - 5 ° S and 150 ° W - 160 ° E.

ENSO takes important rule in extreme variability rain conditions. Fluctuations of ENSO occurance in Pacific Ocean is highly related to rainfall in Indonesia [8]. For global climate variability, NINO 3.4 is more frequently used that has broad impact sea surface temperature variability in this state has thougest effect on rainfall friction in West Pacific [7]. West to Central Pasific friction causes the heating location changes this lead of mostly global atmospheric circulation. Boundary of this research is also appointed so that the scope is not too broad or too narrow. Flowchart of this research method can be seen in Figure 1.

B. Data Preprocessing

Data used are Southern Oscillation Index (SOI), NINO 3.4 SST, and rainfall during dry season precipitation observation data the dry season observe data. SOI and NINO SST data are derivide from Australian Bureau of Meteorology (BOM). SOI data used is from 1876 up to 2012, whereas NINO SST data is from 1950 up to 2010. Rainfall observed data is through 1965 up to 2010 are derivide from *Badan Meteorologi Klimatologi dan Geofisika* (BMKG) through weather station in the district of Indramayu. This research only used May up to February SOI data in 1970 until 2010. So is the case with NINO 3.4 data and the observed data is referred to SOI data time range.

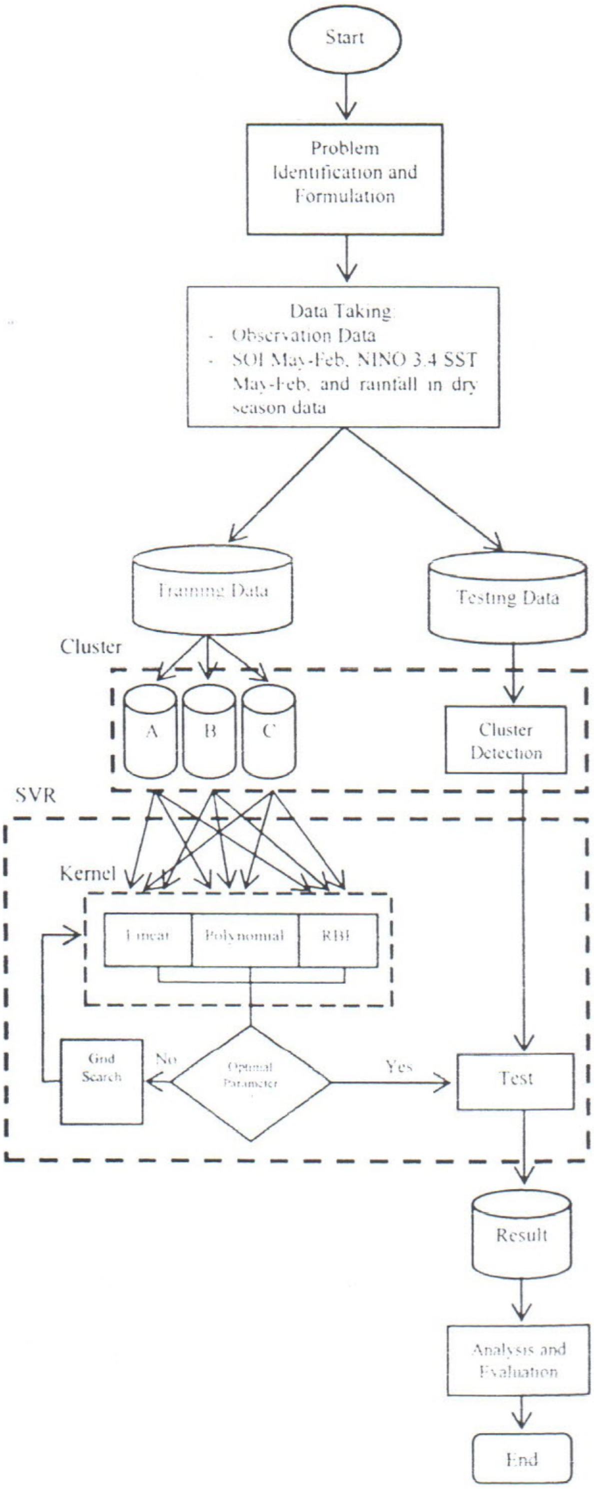

Fig. 1. Flowchart of research processing

While May, June, July, August rainfall during dry season or MJJA RFDS (*Curah Hujan Musim Kemarau Mei Juni Juli Agustus*, CHMK MJJA) data is used as the predicted object. The rainfall data are divided into 5 rainy zones or RaZ (*Wilayah Hujan*, WH) such as:

1. Rainy Zone 1 : Losarang, Pusaka Negara, Sukra, and Ujing Garis
2. Rainy Zone 2 : Sudikampiran and Sudimampir
3. Rainy Zone 3 : Lawang Semut, Teluk Kacang, and Wanguk
4. Rainy Zone 4 : Rentang, Sukadana, and Tugu
5. Rainy Zone 5+6 : Sumurwatu, Taminyang, and Slamet

Rainfall during dry season is obtained from summary of May up to August perdasarian rainfall. Afterwards of RaZ 1 until RaZ 5+6 mean value for Beach rainfall periode years is calculated. SOI and NINO 3.4 data is obtained by using May until February data for each rainfall period years.

Data division aim to find out training data and tested data. Training data is used to form SVR model while tested data is for calculating obtained SVR models accuracy. Tested data used is only applied in a year along.

C. Clustering by Ward Method

Clustering using ward method is implemented in SOI, NINO 3.4 SST, and rainfall during dry season data by random partition in *k* cluster. This method embark the grouping on research units that have similar characteristics are analogous to the closest distance. Next, this training data would be divided into some clusters according to the appointed *k* cluster. After get classes that showed the closest characteristics then cluster detection process is carried on the tested data. Tested data using SVR models corresponding to its cluster.

D. SVR Modelling

SVR is implementation of Support Vector Machine (SVM) for regression case. The output in regression case is real numbers or continuous. SVR is a method that can overcome the overfitting, so it will produce a good performance [9]. The basic idea of Support Vector Regression to determine which data sets are divided into training sets and test sets. Then determined from the training set of a regression function with a certain deviation limits thus produce a prediction that close to the actual target. Training data is processed using SVR training to obtain the model with the data used SOI, NINO3.4 SST, and dry season rainfall data as input for training.

SVR process is implemented in each cluster that has been formed in clustering step. SVR uses kernel functions to transform the non-linear input into the feature space that dimension is higher due to problems in the real world are rarely linear separable generally.

As for some of kernel functions are:

1. Linear Functions

Linear function equation is

$$k(x,y) = x^T y + C$$

2. Polynomial Functions

Polynomial function Equation is

$$k(x,y) = (\alpha x^T y + C)^d$$

3. Gaussian function (RBF)

RBF function equation is

$$k(x,y) = \exp(-\gamma \|x - y\|^2)$$

Kernel function used in SVR are linear, polynomial, and RBF kernel functions. Model performance of kernel function can be known by its correlation coefficient and NRMSE error value. Each kernel function has parameter value that must be appointed firstly. Parameter C value is referred to linear kernel function, parameter C, γ, r, and d value is referred to polynomial kernel function, whereas parameter C and γ is referred to RBF kernel function. The parameter value gives big impact on resulted SVR model. The more optimal parameter, it means the better resulted model. Search of kernel function optimum parameter uses grid search.

E. Test

Cluster detection is implemented to tested data that are SOI and NINO3.4 SST value in one year along. Detection is applied using Squared Pearson calculation to detect cluster in tested data. Next, tested data whose cluster has been known is proceed by SVR model based on its cluster.

In this step, tested data is used as input for SVR models to get of predicted value. The test based on appropriate SVR models with categorized cluster of that tested data.

F. Analysis and Evaluation

Accuracy and eror calculation of predicted results using SVR model to tested data uses correlation coefficient (R) and Normalized Root Mean Square Error (NRMSE). Model compability is considered good if R value is close to 1 and NRMSE is close to 0. Correlation coefficient showed strong relation between two variables. It is described bellow about correlation coefficient R equation:

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{[n\sum_{i=1}^{n} x_i^2 - [\sum_{i=1}^{n} x_i]^2][n\sum_{i=1}^{n} y_i^2 - [\sum_{i=1}^{n} y_i]^2]]}}$$

legend:
x_i : actual value / observed value
y_i : predicted value
n : data amount

Error value is used to determine deviation of estimated value against the actual value. Error calculation uses Normalized Root Mean Square Error (NRMSE). It is described bellow about NRMSE equation:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}}{\sigma_y}$$

legend:
x_i : actual value / observed value
y_i : predicted value
n : data amount

σ_y: deviasi standard of prediction

III. RESULTS AND DISCUSSION

A. Data and Predictors Selection

Rainfall data used in the research is sourced from Indonesia bureu of Meteorology Climatology and Geophysics (*Badan Meteorologi Klimatologi dan Geofisika*, BMKG). More detailed information about the data used can be found in Chapter Methodology. The election of SOI and NINO3.4 SST as a predictor of SVR modeling as it relates to rainfall in dry season. More proper predictors are used, better the resulting model. SOI and NINO is one of indicators of the ENSO phenomenon affecting climate.

Commonly, Indonesia has two seasons, rainy and dry seasons in where rainy season is main factor as most important part Indonesian tropical climate [10]. Other main factors that influence Indonesian climate are monsoon and many other processes like the El Nino Southern Oscillation (ENSO). Global symptoms appearance such as El Nino and La Nina that are caused by ENSO can be predicted by observing anomaly repetitions happened in sea surface temperature.

SOI is anomalies of air pressure difference in the Tahiti surface in Polynesia islands, French, with Darwin surface, Australia. This natural phenomenon is followed by deviations of rainfall circulation and patterns. SOI negative value commonly indicate El Nino phenomenon whereas the positif one which is connected to stronger the Pacific trade winds and warmer sea temperature in north of Australia Mans La Nina phenomenon.

Besides global variable that gives impact on El Nino and La Nina phenomenon is NINO sea surface temperature anomaly. Changes sea surface temperature is closely related to the symptom happened at transformation atmosphere. Symptom is necessary to be observe because of the existance of sea-atmosphere bilateral influence. El Nino and La Nina extreme symptoms is appeared because this interaction by the sea. NINO 3.4 is considered more appropriate to be used than other NINOs. Sea surface temperature variability in this state has the haighest effect on rainfall friction in West Pacific. West-to-Central Pacific friction causes the heating location change thus lead of mostly global atmospheric circulation [7]. So, SOI and NINO 3.4 is used as predictor to predict rainfall during dry season.

Not Ever months of SOI and NINO 3.4 SST's variables would be used as predictor. The months used are May up to February in next year. This election aimed to predict rainfall during dry season of MJJA (May June July) in next year as well.

B. Model Performance Based on SVR Kernel Function

This research used 30 annual training data clustered by k = 3. Clustering and its detection process on tested data were applied by software named MINITAB 16. Each cluster has SVR model with three SVR kernel function, such as Polynomial, Linear, and RBF kernel. Cluster detection on tested data aimed to obtain optimal prediction results by using appropriate SVR models. Annual tested data would use SVR models suited with its cluster. SVR kernel function performance could be seen from its correlation rate and error values of each kernel's obseved data. The model performance is considered good if its correlation rate is high and prediction error value is low.

Training using SVR needs parameter suited with its kernel. To get optimal kernel, grid search implemented in the training is done. Parameter C value is referred to Linear kernel function. Parameter C and γ value is referred to Polynomial and RBF kernel function.

Based on calculation correlation and NRMSE. SVR models with RBF kernel function has highest correlation value and the lowest error, especially correlation value (R) is 0.76 and NRMSE value is 1.73. Best do worst performance model in sequence are RBF, Linear, and Polynomial. Table 1 shows correlation and NRMSE error value of each kernel. Best parameters setting for RBF kernel as the result of grid search is showed by Table 2.

TABLE 1
CORRELATION VALUES AND NRMSE BASED KERNEL

Kernel	Correlation	NRMSE
RBF	0.76	1.73
Linear	0.13	4.37
Polynomial	-0.27	357.54

Every annual tested data ammounted to 10 years became different member cluster. Each cluster of one tested data event has SVR models with dissimilar optimal parameter.

More explanation of kernel function performance in SVR model was described in Figure 2 comparison graph and scatter plots graphs. Comparison graph identified relation between observations (MJJA RFDS) and prediction results of each kernel function. A strong connection between observation and prediction showed more solid correlation and the smaller error between observed and predicted values.

Scatter plot in Figure 3 described relation pattern between observed and predicted values. Linear connection formed straight line indicated that there are close correlation between observed and predicted value. It can be seen that using RBF kernel includes strong connection between observation and prediction result.

TABLE 2
BEST PARAMETERS SETTING FOR RBF KERNEL

Tested Year	Cluster	Best C	Best γ
2000/2001	2	45.25	32
2001/2002	2	2.83	32
2002/2003	2	32	$1.73\text{x}10^{-4}$
2003/2004	3	32	6.10x10-5
2004/2005	2	2	32
2005/2006	1	0.25	0.0625
2006/2007	3	1.41	32
2007/2008	3	33554432	$3.05\text{x}10^{-5}$
2008/2009	2	8	0.13
2009/2010	2	134217728	$4.21\text{x}10^{-8}$

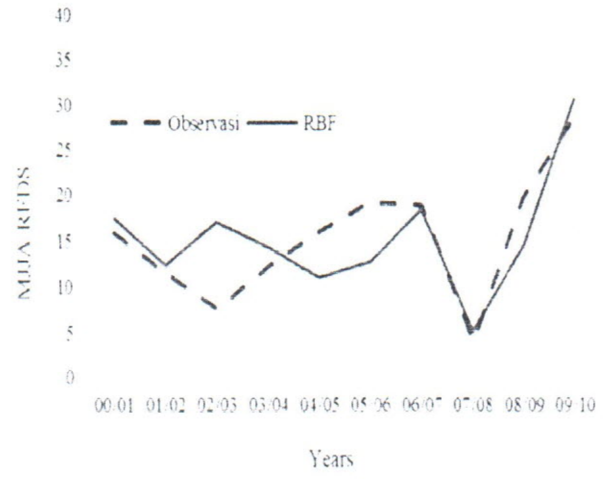

Fig. 2. Comparison graph between observation and prediction based on the RBF kernel performance

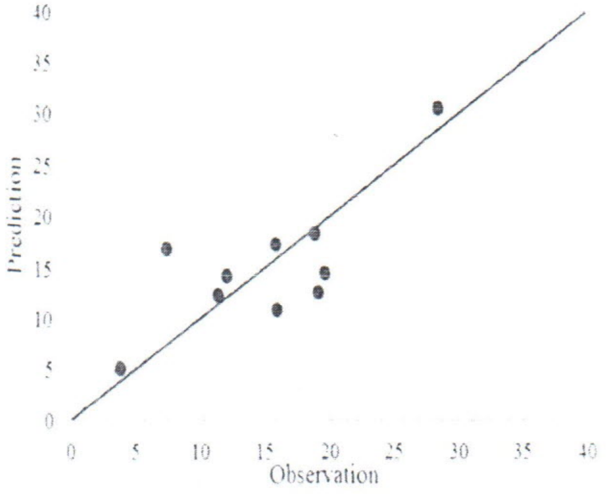

Fig 3. Scatter plot observations with RBF kernel function predictions

C. Analysis and Evaluation Results

Rainfall during dry season prediction using SVR results varied correlation coefficient and NRMSE error values. After training data clustering and tested data cluster detection, it was obtained the best model with the highest correlation rate and the lowest error. Based on SVR best model obtained is using RBF kernel by SOI and NINO 3.4 SST variables in May up to February. NRMSE error value in RBF kernel function is 1.73 as seen in Figure 4. Correlation coefficient value for each kernel function has invers value than its NRMSE error value, as seen on Figure 5.

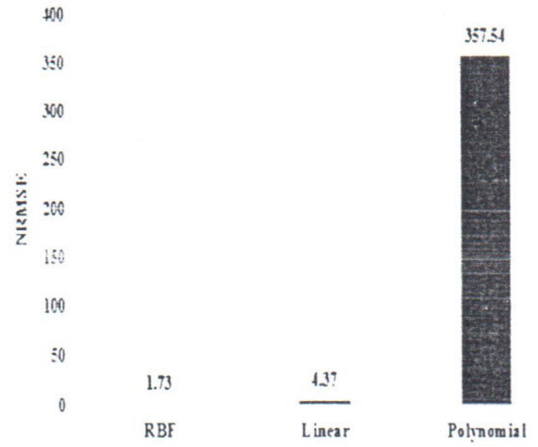

Fig. 4. NRMSE of 3 kernel function

Correlation coefficient RBF kernel is 0.76. It showed that 76% of observation value total variety can be explained by its linear relation with predicted value.

Correlation coefficient Linear kernel is 0.13. It showed that 13% of observation value total variety can be explained by its linear relation with predicted value.

Correlation coefficient Polynomial kernel is -0.27. It explained that negative correlation coefficient has invers connection. It means if the observed value is high, predicted value would be low as well and vice versa. The correlation coefficient value indicates that 26% of observation value total variety can be explained by its linear relation with predicted value.

IV. Conclusion

This research results the best of Support Vector Regression (SVR) model in rainfall during dry season forecasting with highest correlation coefficient value, and lowest NMRSE value using SOI and NINO 3.4 SST data. Tested data using SVR model suited with its cluster to calculate rainfall during dry season prediction value. The SVR model is obtained by using Radial Basis Function (RBF) kernel function and training data cluster amounted to k = 3. Correlation coefficient result gained is 0.76 and NRMSE error value is 1.73. Polynomial kernel function has worst performance by its lowest correlation coefficient and highest NMRSE error values. It is caused by

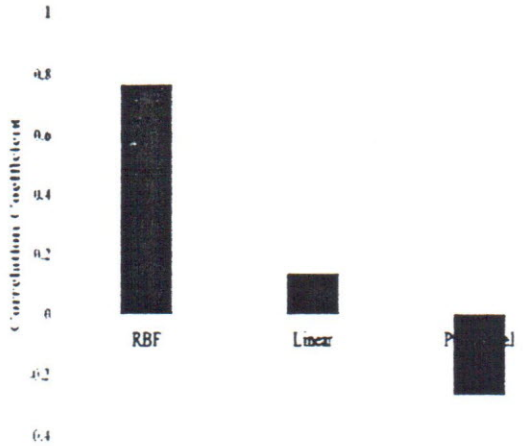

Fig 5 Observation correlation coefficient graph with its prediction using 3 kernel function

incompatibility of function configuration with the data or wrong parameter range election when doing grid search.

ACKNOWLEDGMENT

Gita Adhani thanks to Computer Science Department, Bogor Agricultural University for the supervision; Center for Climate Risk and Opportunity Management in Southeast Asia and Pacific (CCROM-SEAP), BAU, for the supported data; United States Agency for International Development (USAID) for partly supported financial; *Direktorat Jenderal Pendidikan Tinggi* (DIKTI) through the *Bantuan Operasional Perguruan Tinggi Negeri* (BOPTN) Research Program for partly supported financial.

REFERENCES

[1] Estiningtyas W. Wigena A H. 2011. Teknik Statistical Downscaling dengan Regresi Komponen Utama dan Regresi Kuadrat Terkecil Parsial untuk Prediksi Curah Hujan pada Kondisi El Nino. La Nina dan Normal. *Jurnal Meteorologi dan Geofisika*. 12(1):65-72.

[2] [BMKG] Badan Meteorologi Klimatologi dan Geofisika (ID). 2010. Hujan di Musim Kemarau Dampak La Nina. [downloaded 2012 Nov 25]. Available: http://www.bmkg.go.id/RBMKG_Wilayah_9/Lain_Lain/Artikel/HUJAN_DI_MUSIM_KEMARAU_DAMPAK_LA_NINA.bmkg.

[3] Larasati, R. 2012. Prediksi awal musim hujan menggunakan data southern oscillation index dengan metode Support Vector Regression [skripsi]. Bogor (ID): Bogor Agricultural University.

[4] Agmalaro MA. 2011. Pemodelan statistical downscaling data GCM menggunakan support vector regression untuk memprediksi curah hujan bulanan Indramayu [thesis]. Bogor (ID): Bogor Agricultural University.

[5] [BOM] Bureau of Meteorology. 2002. Climate Glossary - Southern Oscillation Index (SOI) [downloaded 2013 Jun 29]. Available: http://reg.bom.gov.au/climate/ glossary/soi.shtml.

[6] As-syakur AR. 2007. Identifikasi hubungan fluktuasi nilai SOI terhadap curah hujan bulanan di kawasan Batukaru-Bedugul, Bali. *J Bumi Lestari*. 7(2):123-129.

[7] [IRI] The International Research Institute for Climate and Society (US). 2007. Monitoring ENSO. [downloaded 2012 Nov 25]. Available: http://iri.columbia.edu/climate/ENSO/background/monitoring.html.

[8] Aldrian, E., L.D. Gates, and F.H.Widodo. 2007. Seasonal variability of Indonesian rainfall in ECHAM4 simulations and in the reanalyses. The role of ENSO. *Theoretical and Applied Climatology*. 87: 41–59 doi:10.1007/s00704-006-0218-8.

[9] Smola AJ. Schölkopf B. 2004. A Tutorial on Support Vector Regression. *Statistics and Computing*. 14:199-222.

[10] Nieuwolt S. McGregor GR. 1982. *Tropical Climatology: An Introduction to The Climate of The Low Latitude*. Chichester (UK): John Wiley anc Sons Ltd