# A Neural Network Architecture for Statistical Downscaling Technique : A Case Study in Indramayu District

Agus Buono[1] , Akhmad Faqih[3], Rizaldi Boer[2], I Putu Santikayasa[4], Arief Ramadhan[5], M. Rafi Muttqien[6], M. Asyhar A[7].

[1,2,3)]Center for Climate Risks and Opportunity Management in Southeast Asia and Pacific (CCROM-SEAP) Bogor Agriculture University, Bogor-West Java, Indonesia

[4)]Department of Geophysics and Meteorology, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor-West Java, Indonesia

[5,6,7)]Department of Computer Sciences, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor-West Java, Indonesia

pudesha@yahoo.co.id, rizaldiboer@gmail.com

**ABSTRACT**

This research is focused on the development of statistical downscaling model using neural network technique to predict SOND rainfall in Indramayu. SST and rainfall data from multimodel ensemble outputs (derived from 18 ensemble members of ECHAM5 model under SRES A1B scenario) is used as predictor to predict SOND rainfall in each station. SST domains were selected by using cluster and correlation analyses, which were divided into three sets, namely SST lag 1 (August), lag 2 (July), and lag 3 (June). The Artificial Neural Network (ANN) employed in this study was multilayer perceptron with hidden layer as many as 5, 10, 20, and 40, and was trained with back propagation. The results show that the observed value lies between the maximum and minimum values of the predicted data. It is shown that the lagged SST provides better relationship with the observed data, and the optimum number of hidden neurons in neural networks is 5. Maximum correlation resulted from the models is 0.796 with an average of about 0.6. It is found that the prediction results tend to overestimate low rainfall and underestimate high rainfall found in the observed data.

Keywords: *General Circular Model (GCM), Statistical Downscaling(SD), Neural Network(NN), Principle Component Analysis (PCA).*

## I. INTRODUCTION

Facts indicated that climatic conditions could contribute significantly to agricultural productions. In this case, many techniques have been developed to predict climate variables that can be used to support agricultural management system. Most of these techniques support the analysis of climatic effects on a particular region.

In developing the prediction models, there are usually two main obstacles, first is the limitation of historical climate data with a sufficiently long series, and second is the need of future climate projections (under certain scenarios) to study the impacts of climate change. General Circulation Model (GCM) provides a solution to this problem and the data has been widely used for climate change studies. However, due to its coarse resolution, that is about 2x2 degrees, or about 200x200 km, the model is unable to capture local variability that is needed in the analysis of a smaller coverage area, such as district level. Therefore there is a gap between the GCM output and the observed data. In this case, the GCM is only able to capture the pattern of average, whereas variability mainly influenced by local factors is not accommodated.

This research is addressed to develop a statistical downscaling model using artificial neural networks (ANN). This model links the rainfall data from GCMs and Sea Surface Temperature (SST) with the observational data to predict rainfall intensity in Indramayu district. Downscaling techniques will be applied to estimate the total rainfall on SOND (September, October, November, and December) season. With 24 time periods of data, an 8-fold cross validation technique is implemented to evaluate the model.

The remainder of this paper is organized as follows: Section 2 presents the principles of statistical downscaling. Section 3 describes the methodology used in this study Result and discussion is presented in Section 5, and finally, Section 6 is addressed to the conclusions of this research study.

## II. STATISTICAL DOWNSCALING

Downscaling is defined as an effort to connect between global-scale (explanatory variables) and local scale climate variables (response variables), [1]. Figure 1 illustrates the process of downscaling.

There are two approaches for downscaling, using regional data (obtained from a regional climate model, RCM), or global data (obtained from the general circulation models, GCM). The first approach is known as statistical dynamical downscaling, while the second is known as statistical downscaling (SD). Statistical downscaling based on the relationship between coarse-scale grid (predictor) with local-scale data (response) is expressed with a statistical model that can be used to translate a global scale anomalies which became an anomaly of some variables of local climate (Zorita and Storch 1999, in [2]). In this case the SD is a transfer function that describes the functional relationship of global atmospheric circulation with elements of the local climate, which is formulated as follows:

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

where :
Y : response climate variables
X : global climate variables (provided by GCM)
t : time period
p : dimension of Y
q : dimension of X
s : layers in the atmosphere
g : GCM domain

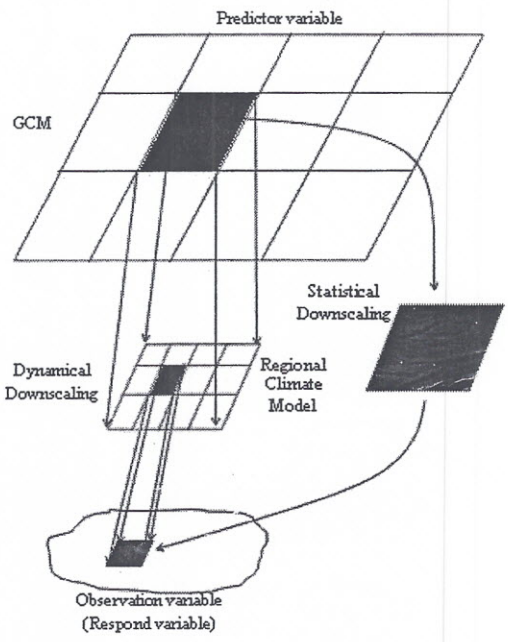



Figure 1. Illustration of the downscaling (Source : [3])

In general SD model involving time series data (t) and spatial data of GCM (g). Number of Y, X variables, the layer of the atmosphere in the model and the autocorrelation and co-linearity on the variables Y and X indicate the complexity of the model. Until now the SD models that have been developed are generally categorized into five, i.e., i) based on regression techniques or classification, ii) based on linear or non linear model, iii) based on parametric and non parametric, iv) based on projection and selection, and v) based on model-driven or data-driven techniques. Nevertheless, an SD model can be included in the combination of the five categories, for example PCR (principle component regression) that were categorized as regression-based methods, linear, parametric, projections and data-driven. In this research, we developed an adaptive neural network (ANN) model for statistical downscaling using data from the GCM and sea surface temperature as explanatory variables. The use of SST data is specifically intended to capture the El Nino phenomenon, so the model is expected to produce better prediction results.

## III. DATA AND EXPERIMENTAL SETUP

### A. Data

The research involved three types of data, i.e. i) precipitation data from GCM model (with the A1B scenario, ECHAM5 model (with 18 members and 2.8°x2.8° resolution), ii) SST data (with 2°x2° resolution), and iii) rainfall data from 17 rain gauge stations in Indramayu. All datasets have the time period from 1979 to 2002 (24 periods). Figure 2 give the spatial distribution of the stations in Indramayu District.

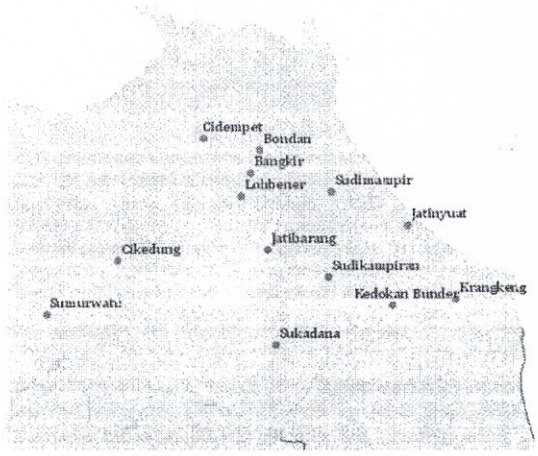



Figure 2. Spatial distribution of the climate station in Indramayu

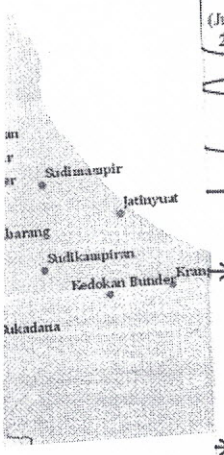
### B. Experimental Setup

Figure 3 shows five stages of the experimental setup used in this study, i.e.:

a. Preprocessing : This process consists of (1) checking the rainfall observation data (validity and consistency) using the method described by [4]; (2) calculating the SOND season climate variables (rainfall observation of all stations, GCM for all ensemble members). Especially for the SST, the data is divided into three months, i.e. June (lag 3), July (lag 2) and August (lag 1); and (3) calculating the normalization of SOND data for rainfall observations and GCM data, and the normalization of SST for each lag.
b. SST domain selection: clustering the SST domain into N clusters, and checking the correlation between the normalized observation rainfall with the center of the SST cluster. In this case, the SST data in particular grid is taken as part of the SST domain set if the correlation is found to be statistically different from zero at 0.9 confidence level.
c. Feature Extraction: For each rainfall station, the GCM domain that is used as predictor is within the dimension of

## NTAL SETUP

types of data, i.e.
l (with the A1B scena
embers and 2.8°x
2°x2° resolution),
ge stations in Indram
od from 1979 to 2
spatial distribution of




on of the climate station
mayu

of the experimental setu

ess consists of (1) checki
validity and consistency)
[4]; (2) calculating the
(rainfall observation
emble members). Especia
led into three months, i.
and August (lag 1); a
ation of SOND data for
data, and the normaliza

lustering the SST domain
the correlation betwe
rainfall with the center of
e SST data in particular
T domain set if the corre
lly different from zero

r each rainfall station, th
redictor is within the dime

5. After that, 25 dimensional vector GCM is reduced by ing principle component analysis.

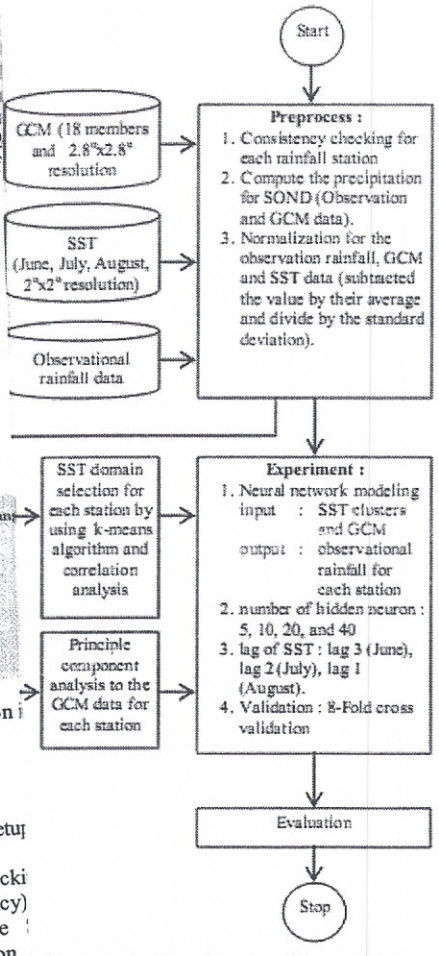



Figure 3. The flow diagram of the experiments

deling and testing ANN: In this ANN model, inputs are GCM and SST data, while the outputs are rainfall data of all stations. The developed ANN el is the multilayer perceptron with one input layer, hidden layer and one output layer. Input layer sists of two groups of neurons, i.e. one group for and another group for rainfall from GCMs. The ut layer is in accordance with a consistent number tations. Training of the ANN model is based on error propagation algorithm as described by [5]. Two idered factors in this experiment are the number of en neurons (i.e.: 5, 10, 20, and 40), and the lag time e SST data, i.e. the SST in June (lag 3), July (lag 2), August (lag 1). By considering that the total period ta is 24, the 8-fold cross validation is then used to he model,

e. Model evaluation: the ability of neural network techniques in predicting the total rainfall was evaluated by comparing the predicted data with observations.

## IV. RESULT AND DISCUSSION

Consistency analysis on 17 rainfall stations in Indramayu show four stations that are inconsistent. Those stations are Losarang, Indramayu, Bulak Kandanghaur and Tugu. These are indicated by the values of F statistics that are greater than the F threshold (4.3248) at 95% confidence level as shown in Figure 4.

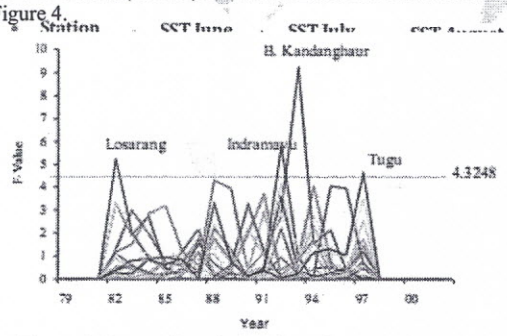



Figure 4. Pattern F statistic values for 17 stations.

Clearer illustration of the inconsistencies is shown in Figure 5. The figure shows that the SOND rainfall inconsistent in 1993-1994. Based on the results above, the four stations are not included for further analysis, which mean that only 13 stations will be analyzed in this study.

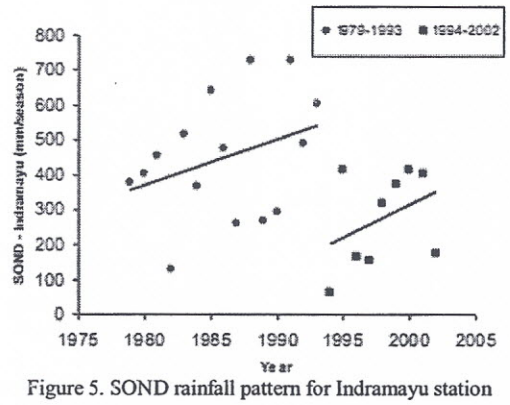



Figure 5. SOND rainfall pattern for Indramayu station

The result of domain selection for SST (using clustering techniques and correlation analysis) shows that the SST record in August (lag one) have a higher correlation than the lag 2 and lag 3, as shown in Figure 6. The average correlation between the observations and SST for lag 1, lag 2 and lag 3 are 0.443, 0.461, and 0.483, respectively. While the maximum correlations are 0.763, 0.7184 and 0.7964, each for lag 1, lag 2 and lag 3, respectively.

5x5. After that, 25 dimensional vector GCM is reduced by using principle component analysis.

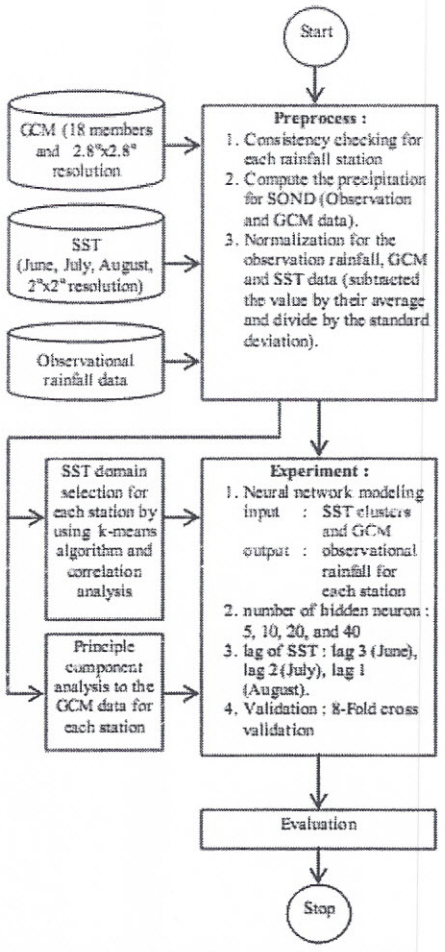



Figure 3. The flow diagram of the experiments

d. Modeling and testing ANN: In this ANN model, the inputs are GCM and SST data, while the outputs are the rainfall data of all stations. The developed ANN model is the multilayer perceptron with one input layer, one hidden layer and one output layer. Input layer consists of two groups of neurons, i.e. one group for SST and another group for rainfall from GCMs. The output layer is in accordance with a consistent number of stations. Training of the ANN model is based on error back propagation algorithm as described by [5]. Two considered factors in this experiment are the number of hidden neurons (i.e.: 5, 10, 20, and 40), and the lag time of the SST data, i.e. the SST in June (lag 3), July (lag 2), and August (lag 1). By considering that the total period of data is 24, the 8-fold cross validation is then used to test the model,

e. Model evaluation: the ability of neural network techniques in predicting the total rainfall was evaluated by comparing the predicted data with observations.

## IV. RESULT AND DISCUSSION

Consistency analysis on 17 rainfall stations in Indramayu show four stations that are inconsistent. Those stations are Losarang, Indramayu, Bulak, Kandanghaur, and Tugu. These are indicated by the values of F statistics that are greater than the F threshold (4.3248) at 95% confidence level as shown in Figure 4.

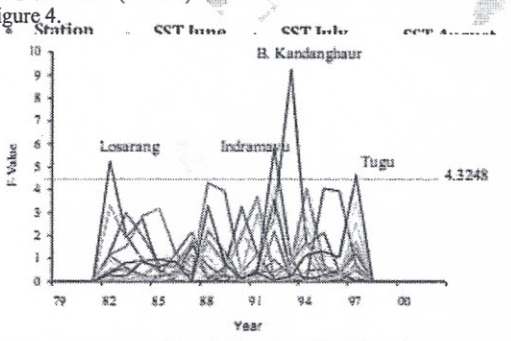



Figure 4. Pattern F statistic values for 17 stations.

Clearer illustration of the inconsistencies is shown in Figure 5. The figure shows that the SOND rainfall inconsistent in 1993-1994. Based on the results above, the four stations are not included for further analysis, which mean that only 13 stations will be analyzed in this study.

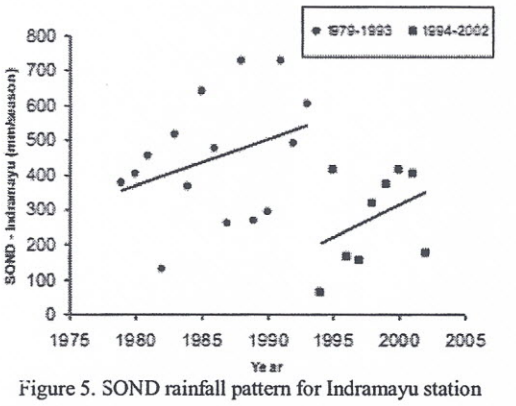



Figure 5. SOND rainfall pattern for Indramayu station

The result of domain selection for SST (using clustering techniques and correlation analysis) shows that the SST record in August (lag one) have a higher correlation than the lag 2 and lag 3, as shown in Figure 6. The average correlation between the observations and SST for lag 1, lag 2 and lag 3 are 0.443, 0.461, and 0.483, respectively. While the maximum correlations are 0.763, 0.7184 and 0.7964, each for lag 1, lag 2 and lag 3, respectively.

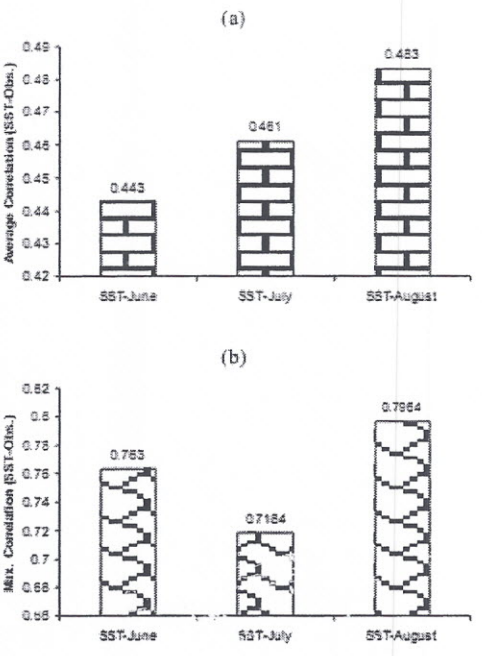



Figure 6. Correlation between SST with SST observations for different lag (a: average correlation, b: maximum)

Figure 7 shows the differences in SST domain that provides the dominant influence on the observation at particular rainfall stations (the figure only presents for station Sumurwatu and station Cikedung). From the figure we can see that for different stations, the SST domain is also different. In a particular station, the SST domain is also different, if the lag time for SST is also different. The differences in SST domain occur both spatially and temporally.

In accordance with the above analysis, the neural network architecture constructed in this research is shown in Figure 8. The figure shows that the number of output neurons is 13, i.e. the number of stations that are consistent, the number of hidden neurons are tested are 5, 10, 20, and 40. While the number of input neurons is in accordance with the number of SST clusters that correlated significantly with observations coupled with the dimension of GCM resulted by the principal component analysis. Then subsequently carried out the training and testing of the neural network models by following the scenario 8-fold cross validation. The model is trained using back propagation algorithm.

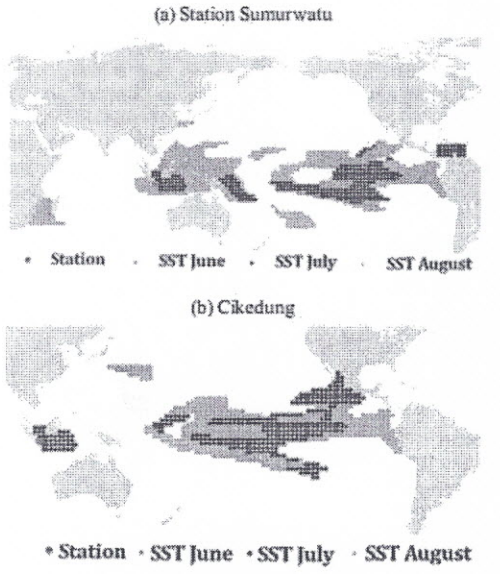



Figure 7. Domain distribution of SST (a: Sumurwatu; b: Cikedung)

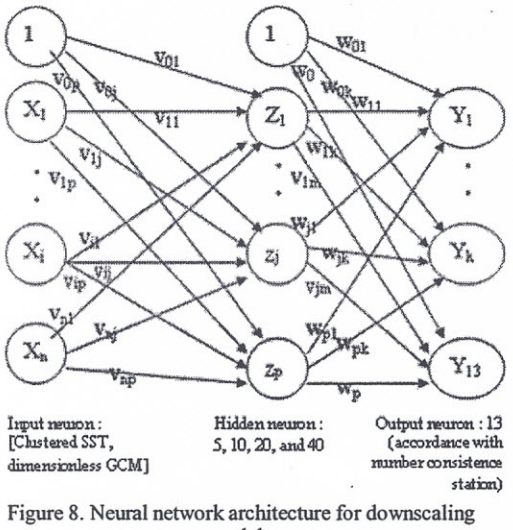



Figure 8. Neural network architecture for downscaling models

Figure 9 presents the boxlot for the observation and the predicted value for different number of hidden neurons and the various SST lag. From the pictures can be seen that the greater the number of hidden neurons, the more extreme predicted values appear. From the picture we can conclude that the appropriate number of hidden neurons is 5.

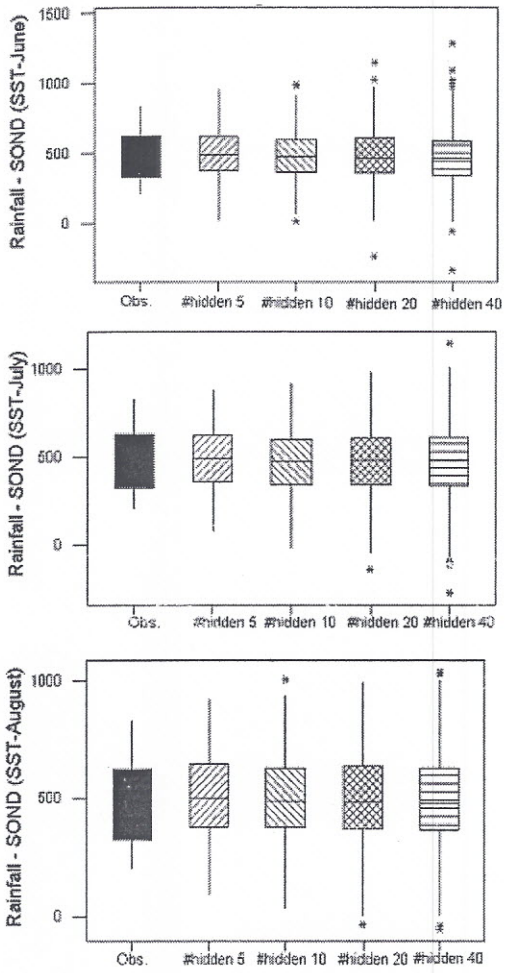



Figure 9. Boxplot for observation and the predictive value at different number of hidden neurons (a: SST lag 3, b: SST lag 2, c: SST lag 1)

Figure 10 presents the pattern of SOND rainfall for the observation (blue), the average predicted value from the 18 members of the GCM (green), maximum predicted value (black) and the minimum of the predicted value (red). From the pictures it can be seen that the rainfall observations are generally located between the maximum and minimum values of the prediction. In year 1992, it was shown that the observation is above the maximum of the predicted value

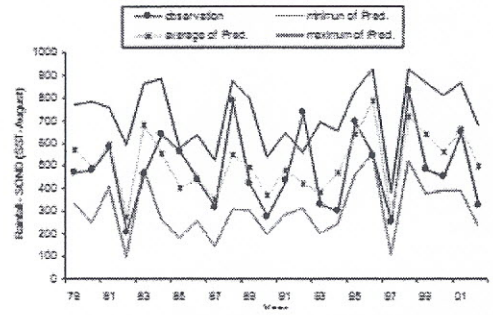



Figure 10. Comparisons between the SOND rainfall pattern for the observations and their prediction (average, minimum and the maximum)

The comparisons between observations and average predictions (resulted by averaging the prediction from 18 members of the GCM) for the SST lag 1, lag 2 and lag 3 are presented in Figure 11.

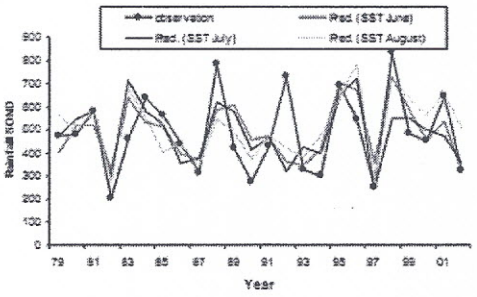



Figure 11. Comparison of observations with an average of predictions for different lag

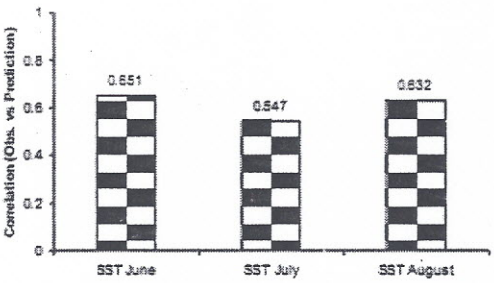



Figure 12. Comparison of correlation between the observations and the average prediction for different lag

As presented in the Figure 11, it can be seen that the overall pattern according to the pattern predicted observation. An extreme deviation occurred in 1988 and 1992. In 1987 up to 1995 shows that the downscaling technique is not able to follow the extremes pattern that exist in the data. Correlation between observations and predictions range from 0.547 to 0.651 as shown in Figure 12.

Figure 13 presents the scattered plots between observations and predictions. The figure demonstrates that for small values of observations, the predicted value tend to overestimate, and for the higher value of the observation that there tends to be underestimated.

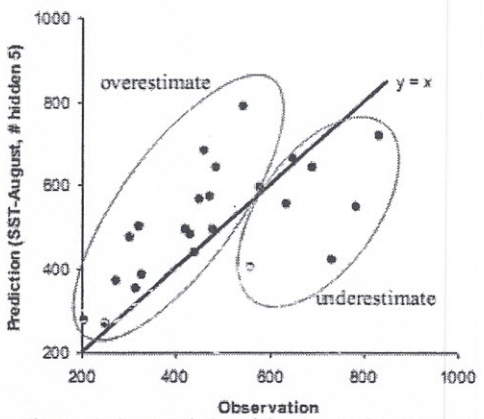



Figure 13. Scattered plot of the observation vs. prediction

## V. CONCLUSSION

Based on the experiment, we can conclude several things:

a. SST domain influencing the observations at a particular station varies temporally and spatially. In this case, different stations have different SST domain. At a station, if the lag is different, the domain in SST is also different.
b. SST with a lag (SST in August) gave the highest correlation with the observed data that is equal to 0.796. While the correlations for the lag 2 and lag 3 are 0.718 and 0.763, respectively.
c. ANN with hidden neuron 5 is capable for predicting SOND rainfall with correlation between observations and predictions, i.e. 0.651, 0.547, and 0.632 for the SST lag 3, lag 2 and lag 1, respectively.
d. There is a tendency that the models overestimate low SOND rainfall and underestimate high SOND rainfall found in the observed data.

## ACKNOWLEDGMENT

This research is part of Bogor Agricultural University's IM-HERE B2c Project, which is funded by the National Government under contract number: 7/13.24.4/SPP/I- MHERE/2010. The authors thank to International Research Institute (IRI), Columbia University for their permission to access the GCM datasets and to BMKG for providing the rainfall data.

## REFERENCES

[1] Wilby, R.L. and T.M.L. Wigly. Downscaling General Circulation Model Output : a Review of Methods and Limitations. Progress in Phisical Geography 21, 4, pp. 530-548, 1997.

[2] Wigena, A.H. Pemodelan Statistical Downscaling dengan Regresi Projection Pursuit untuk Peramalan Curah Hujan Bulanan : Kasus Curah Hujan Bulanan di Indramayu. PhD Dissertation, Department of Statistics, Graduate School, Bogor Agriculture University, Bogor 2006.

[3] Sutikno. Statistical Downscaling Luaran GCM dan Pemanfaatannya untuk Peramalan Produksi Padi. PhD Dissertation, Department of Meteorology, Graduate School, Bogor Agriculture University, Bogor. 2008.

[4] Boer, R. Metode untuk mengevaluasi Keandalan model Prakiraan Musim. Climatogy laboratory, Departement of Meteorology, Facultu of Mathematics and Natural Sciences, Bogor Agriculture University Bogor. 2006.

[5] Laurene F. Fundamentals of Neural Networks. Prentice Hall, Englewood Cliffs. 1994.