ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 22, NO. 5, 2005, 703–710

Statistical Prediction of Heavy Rain in South Korea

Keon Tae SOHN*[1], Jeong Hyeong LEE[2], Soon Hwan LEE[3], and Chan Su RYU[3]

[1] *Pusan National University, Busan* 609-735, *Korea*

[2] *Dong-A University, Busan* 604-714, *Korea*

[3] *Chosun University, Gwangju* 501-759, *Korea*

(Received 18 January 2005; revised 22 April 2005)

ABSTRACT

This study is aimed at the development of a statistical model for forecasting heavy rain in South Korea. For the 3-hour weather forecast system, the 10 km×10 km area-mean amount of rainfall at 6 stations (Seoul, Daejeon, Gangreung, Gwangju, Busan, and Jeju) in South Korea are used. And the corresponding 45 synoptic factors generated by the numerical model are used as potential predictors. Four statistical forecast models (linear regression model, logistic regression model, neural network model and decision tree model) for the occurrence of heavy rain are based on the model output statistics (MOS) method. They are separately estimated by the same training data. The thresholds are considered to forecast the occurrence of heavy rain because the distribution of estimated values that are generated by each model is too skewed. The results of four models are compared via Heidke skill scores. As a result, the logistic regression model is recommended.

Key words: heavy rain, model output statistics, linear regression, logistic regression, neural networks, decision tree

1. Introduction

In recent decades, extreme weather events seem to be growing in frequency and risk due to water-related disasters. According to the World Meteorological Organization report (ISDR and WMO, 2004) on World Water Day, 22 March 2004, the economic losses caused by water-related disasters, including floods, droughts and tropical cyclones, are on an increasing trend as follows: the yearly mean in the 1970s was about 131 billion US dollars, 204 billion dollars in the 1980s, and 629 billion dollars in the 1990s. Almost two billion people were affected by natural disasters in the last decade of the 20th century, 86% of them by floods and droughts.

In South Korea, economic losses caused by natural disasters are also increasing as follows: the yearly mean in the 1970s was about 0.135 billion dollars, 0.363 billion dollars in the 1980s, and 0.530 billion dollars in 1990s. Table 1 shows that yearly losses caused by water-related disasters have an increasing trend during the last 10 years (1993–2002). And the mean was 0.24 percent GDP. In 2002, economic loss caused by the super typhoon Rusa was over 5 billion dollars. At that time, a new record for daily maximum amount of rainfall, 870.5 mm d^{-1}, was made in Gangneung city. In 2003, economic loss caused by the typhoon Maemi was over 4 billion dollars. Table 2 shows that typhoons and heavy rainfall in the summer season are the main disastrous weather events with a large loss of life and property.

The improvement of the prediction of heavy rainfall is very important in order to reduce its potential for damage. There are four types of forecasts: (1) the

Table 1. Yearly losses caused by water-related disasters in South Korea (KMA*).

Year	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Property damage (million $)	216	164	614	478	182	1343	1057	548	1049	5096

*KMA: Korea Meterological Administration.

*E-mail: ktsohn@pusan.ac.kr

Table 2. Sources of property damage caused by water-related disasters in South Korea from 1999 to 2002 (KMA).

Source	Typhoons	Heavy rainfall	Heavy rainfall and typhoons	Heavy snow	Storms	Other
Percent(%)	46.3	30.7	14.7	7.6	0.3	0.4

dichotomous forecast (whether we will have a heavy rain or not), (2) the probability of heavy rain (what is the probability that we will have heavy rain?), (3) the probabilities of classified precipitation (what are the probabilities that the amount of rainfall will belong to the given categories?), and (4) the quantitative precipitation forecast (QPF). This study focuses on the dichotomous forecast of the occurrence of heavy rain. That is, forecasts have binary values: whether we will have heavy rain or not.

Four statistical models (linear regression model, logistic regression model, neural network model and decision tree model) are separately applied to predict the occurrence of heavy rain. And the statistical modeling is performed via the model output statistics (MOS) method. The MOS method, proposed by Glahn and Lowry (1972), is a physical/statistical modeling technique for finding the statistical relationship between numerical model outputs and observations. Many authors have considered the MOS method for the prediction of temperature and precipitation (Lemcke and Kruizinga, 1988; Ross, 1989; Kok and Kruizinga, 1992; Sohn and Kim, 2003).

It is well known that the logistic regression model is useful to the binary response (Myers et al., 2002). As a decision-making problem, the determination of the threshold is required for generating forecasts. That is, a forecaster says that heavy rain will occur when the estimated value is greater than the threshold which is determined based on the distribution of estimated probabilities of heavy rain. The 2×2 (observations×forecasts) tables and receiver operating characteristic (ROC) curves are used in order to determine the threshold. The ROC curve is useful to dichotomous decision-making (Raubertas et al., 1994; DeNeef and Kent, 1993).

And, the neural network model and the decision tree model are also considered because of their nonlinearity. Kim et al. (2001) applied a neural network to the long-range forecast of precipitation in the Seoul area. The classification and regression tree (CART: Breiman et al., 1984), is a tree-based method for the classification and prediction. The CART partitions the feature space into a set of rectangles and fits a constant (for instance, mean) in each one. The CART algorithm consists of the selection of splits, the determination of terminal nodes and the estimation rule.

The dataset and potential predictors for our study are presented in section 2. And four models are separately estimated by the same training data and checked by the same validation data in section 3. For our study, the statistical package called SAS, specially SAS E-miner, is used to estimate the parameters of the models and to generate the ROC curves and the frequency tables for training data and validation data. The comparison of three models is presented via the Heidke skill scores (HSS, Heidke, 1926) in section 4. By eliminating the effect of a random forecast, the HSS can be a useful measure to compare forecast models that generate categorical forecasts.

2. Data

The 3-hour weather forecast system is now running at the KMA. For our study, 3-hour-interval data during 2000 to 2003 at 6 stations (Seoul, Daejeon, Gangreung, Gwangju, Busan, Jeju) in South Korea are used. The observations are 10 km×10 km area-mean amounts of rainfall. The predictand, obtained from observation, has a binary response (whether the heavy rain occurred or not). In the KMA, it is defined that heavy rain occurs when the rainfall is over 80 mm d^{-1} or over 10 mm h^{-1}. It is defined by being over 20 mm in 3 hours for this study on the 3-hour weather forecast system.

The 45 synoptic factors at 6 stations are used as potential predictors. These factors include the wind direction and speed, relative vorticity, humidity, thermal advection, potential precipitation and temperatures. They can be generated by the numerical model, called RDAPS (Regional Data Assimilation and Prediction) used in the KMA. And some previous observations are added. Table 3 summarizes them. Cho and Choi (1995) used these predictors for the probability of precipitation and Sohn and Kim (2003) also used these factors for modeling a statistical prediction of precipitation in the Seoul area.

According to Cho and Choi (1995), the climatic characteristics of the warm season (April to October) are different from those of the cold season (November to March) in Korea. So we consider only the warm season. The model training data period is the warm season during 2000 to 2002. And the model validation data period is the warm season in 2003. There are

Table 3. Potential Predictors.

Symbol	Predictors
E850, E700, E500	East wind speed at 850 hPa, 700 hPa and 500 hPa
S850, SE700, S500	South wind speed at 850 hPa, 700 hPa and 500 hPa
NW850, NW700, NW500	Northwest wind speed at 850 hPa, 700 hPa and 500 hPa
NE850, NE700, NE500	Northeast wind speed at 850 hPa, 700 hPa and 500 hPa
VV850, VV700, VV500	Wind speed at 850 hPa, 700 hPa and 500 hPa
VOR850, VOR700, VOR500	Relative vorticity at 850 hPa, 700 hPa and 500 hPa
QAD850, QAD700	Advection of specific humidity at 850 hPa and 700 hPa
Q84	Difference between specific humidity at 850 hPa and 700 hPa
Q74	Difference between specific humidity at 700 hPa and 700 hPa
TAD850, TAD700	Thermal advection at 850 hPa and 700 hPa
RH850, RH700, RH500	RH at 850 hPa, 700 hPa and 500 hPa
CCL	Convective condensation level
DWL	Depth of wet level
PCWT	Potential precipitation
CTOP	Level of cloud top
CBAS	Level of cloud base ’
BBX1	Black box index 1
BBX2	Black box index 2
SSI	Showalt stability index
KYID	KY index
KIDX	K index
LR87	Lapse rate between 850 hPa and 700 hPa
LR85	Lapse rate between 850 hPa and 500 hPa
T850, T700, T500	Temperature at 850 hPa, 700 hPa and 500 hPa
ET850, ET700	Equivalent potential temperature at 850 hPa and 700 hPa
ET87	Difference between equivalent potential temperature at 850 hPa and 700 hPa
Rain3	Drain63 Observed rainfall amount before 3 hours
Drain63	Rain3–Rain6

17928 cases in the training data and 5862 cases in the validation data.

3. Statistical prediction of heavy rain

3.1 *Logistic regression model*

It is well known that the logistic regression model is useful in the binary response case (e.g., “heavy rain”=1, “no heavy rain”=0). The logistic regression model as a generalized linear model (Myers et al., 2002) is defined by the following three components.

(1) The response is a Bernoulli random variable Y where $Y = 0$ if no heavy rain, $Y = 1$ if heavy rain. Let p be the probability of heavy rian. That is $p = P(Y = 1\,|\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector of observed predictors.

(2) The linear predictor is $\boldsymbol{x'\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$.

(3) The link function is a logit link which is defined by logit(p)=log$[p/(1-p)] = \boldsymbol{x'\beta}$. And then $E(Y) = p = [1 + \exp(-x'\beta)]^{-1}$ has an S-shaped function like Fig. 1.

The procedure of the logistic regression modeling is as follows.

(1) Variable selection step: The optimal significant predictors are selected by the stepwise regression method using the training data.

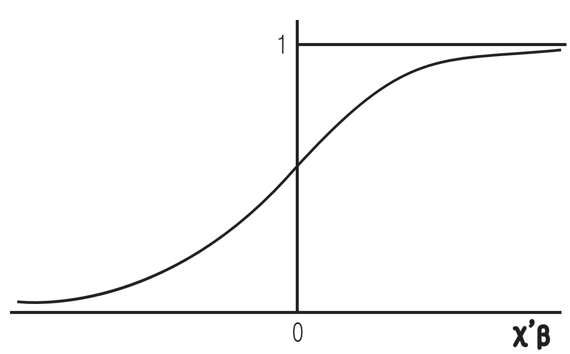

Fig. 1. Shape of the logistic function.

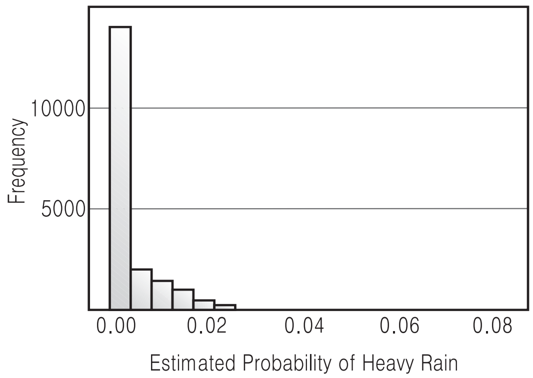

Fig. 2. Histogram of p, the probability of heavy rain.

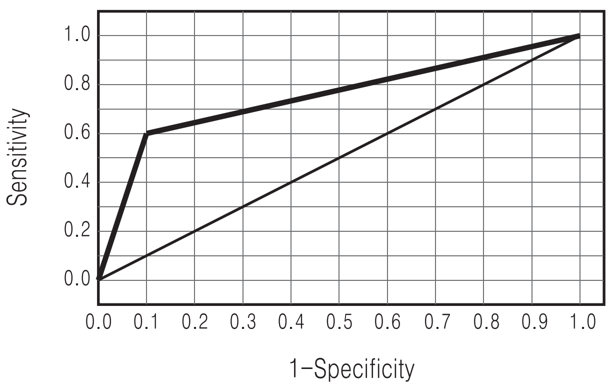

Fig. 3. Receiver Operating Characteristic (ROC) curve.

Table 4. 2×2 table.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(0)	A	B
Heavy rain(1)	C	D

Table 5. 2×2 table for the logistic regression model training.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(17809 cases)	16046(90.10%)	1763(9.90%)
Heavy rain(119 cases)	17(14.29%)	102(85.71%)

Table 6. 2×2 table for the logistic regression model validation.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(5775 cases)	5347(92.59%)	428(7.41%)
Heavy rain(87 cases)	16(18.39%)	71(81.61%)

(2) Estimation step: Parameters in the final logistic regression model are estimated.

(3) Determination of threshold: With the distribution of p (the probability of heavy rain), the threshold of p is chosen as the criterion of the occurrence of heavy rain.

(4) Model validation step: Using the estimated equation and threshold, the model validation is performed and the results of the model validation are compared to those of the model training.

Using the above procedure, the estimated logistic regression model is given by

$$\hat{p} = \frac{1}{1 + \exp(-x'\hat{\beta})} \,, \tag{1}$$

where

$$\begin{aligned} \boldsymbol{x'\beta} =& 61.1749 + 0.00171 \times \text{PCWT} + 0.00243 \times \text{rain3} \\ &+ 0.0605 \times \text{S500} - 0.2844 \times \text{T500} \\ &+ 0.00431 \times \text{VOR700} \,. \end{aligned}$$

A threshold T is needed in order to forecast whether the heavy rain will happen or not. The threshold should be determined based on the distribution of $\hat{p}$. That is, the weather forecast will be "Heavy rain will occur" if the estimated value of the predictand is greater than or equal to T. Since the distribution of p is very right-skewed like Fig. 2, we consider the 2×2 table (Table 4) and the Receiver Operating Characteristic (ROC) curve.

The ROC curve, varying the value of threshold from 0 to 1, is a plot of $(1 - S_1, S_2)$ where $S_1 = A/(A+B)$ and $S_2 = D/(C+D)$. That is, $(1 - S_1)$ is the misclassification rate when the observation is 0 (no heavy rain), and S_2 is the correction rate when the observation is 1 (heavy rain). Here, S_1 means specificity and S_2 means sensitivity

In Fig. 3, the ROC curve of the estimated logistic regression model, there is one change point nearby the point (0.1, 0.6). And then we decided that the threshold T is determined to be 0.01 subject to the following condition:

$$1 - S_1 = P(\text{Forecast} = 1 | \text{Observation}) = 0 \approx 0.1 \,. \tag{2}$$

That is, the misclassification rate is about 0.1 when the observation is equal to 0. By varying the value of threshold T, 2×2 frequency tables are generated until the value of $(1 - S_1)$ is equal to 0.1. In this case, the value of T is determined by 0.011. Table 5 and Table 6 show the results of model training and validation based on the determined value of T. It is a reasonable result that the components of the two tables are simi-

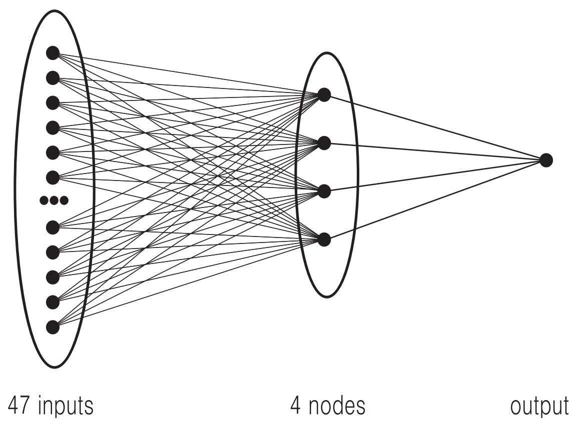

Fig. 4. Structure of the neural network.

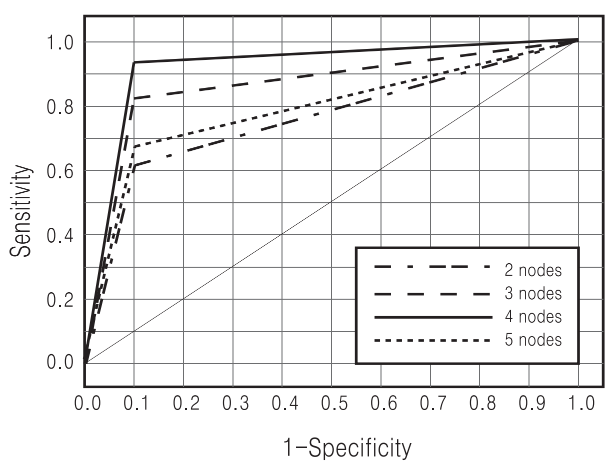

Fig. 5. ROC curves for different numbers of nodes.

Table 7. 2×2 table for the neural network model training.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(17809 cases)	16123(90.53%)	1686(9.47%)
Heavy rain(119 cases)	4(3.36%)	115(96.64%)

Table 8. 2×2 table for the neural network model validation.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain (5775 cases)	5403(93.56%)	372(6.44%)
Heavy rain(87 cases)	45(51.72%)	42(48.28%)

lar to each other.

3.2 *Neural network model*

A neural network model is applied as a supervised statistical learning with many components: 47 inputs, one hidden layer with 4 nodes, and one output (Fig. 4). A linear basis function is used as a combination function and a logistic function is used as an activation function. Optimal weights are estimated via the back propagation algorithm (Haykin, 1999). The ROC can be used as a criterion of model selection. It is reasonable that the model which has the smallest values of $(1-S_1)$ and the largest values of Sensitivity may be the best model. According to the definition of the ROC curve, we consider the neural network with 4 nodes in the hidden layer, which has the greatest ROC curve in Fig. 5.

The distribution of the estimated predictand $\hat{p}$ is more right-skewed than that of the logistic regression model. In the ROC curve of the estimated neural network model, there is also one change point nearby the point (0.1, 0.9). And then we decided that the threshold T is determined to be 0.00165 subject to the same condition: $1-S_1 \approx 0.1$. Table 7 shows that the results of model training based on the determined threshold are very good. Table 8 however says that the results of the model validation are not good.

3.3 *Decision tree model*

The decision tree, called the classification and regression tree (CART), is the tree-based method proposed by Breiman et al. (1984). The CART partitions the feature space into a set of rectangles and fits a constant (for instance, mean) in each one. Let $\{(x_i, y_i)\}_{i=1}^N$ be the training data where x_i is the vector of predictors at the ith observation and y_i is the target value at the same observation. With the training data, the CART constructs a binary tree by proceeding as follows.

(1) Selection of splits: Let a node t denote a subset of the current tree T. Let $N(t)$ denote the total number of cases in node t.

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{x_i \in t} y_i \ , \quad E(t) = \frac{1}{N(t)} \sum_{x_i \in t} [y_i - \bar{y}(t)]^2 \ ,$$

and

$$E(T) = \sum_{t \in T} E(t) \ . \tag{3}$$

Given any set of splits S of a current node t in T, the best split s^* is that split in s that most decrease $E(T)$. Suppose that for any split s of node t into t_L (a new node to the left of t) and t_R (another new node to the right of t). The best split s^* is taken to be the particular split for which we have.

(node t into t_L and t_R) of a current node t in T, the best split s^* is taken to be the particular one for which we have

$$\Delta E(s^*, t) = \max_{s \in S} \Delta E(s, t) \ , \tag{4}$$

Table 9. 2×2 table for the CART model training.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(17809 cases)	15467(86.85%)	2342(13.15%)
Heavy rain(119 cases)	10(8.40%)	109(91.60%)

Table 10. 2×2 table for the CART model validation.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(5775 cases)	5075(87.88%)	700(12.12%)
Heavy rain(87 cases)	8(9.20%)	79(90.80%)

where

$$\Delta E(s,t) = E(T) - E(t_{\rm L}) - E(t_{\rm R}) \ .$$

(2) Determination of a terminal node: A node t becomes a terminal node if the condition $\max_{s\in S}\Delta E(s,t) < \beta$ is satisfied where β is a prescribed threshold.

The distribution of p and the ROC curve are very similar to those of the logistic regression model. The change point of the ROC curve is nearby the point (0.1, 0.55). The threshold is also determined to be 0.01 subject to $1 - S_1 \approx 0.1$. Table 9 and Table 10 show the results of model training and validation respectively. Figure 6 is the final tree which consists of Rain3, DWL, PCWT, S850, E850 and RH850.

3.4 *Multiple linear regression model*

The multiple linear regression model is also applied. The estimated equation is given by

$$\begin{aligned}\hat{p}=&0.61074+0.00039113\times\text{rain3}+0.00708\times\text{KYID}\\&+0.00009688\times\text{VOR850}+0.00010619\times\text{drain63}\\&+8.907452\times10^{-7}\times\text{BBX1}-0.00036927\times\text{E700}\\&-1.17932\times10^{-7}\times\text{BBX2}-0.00010282\times\text{RH850}\\&+0.00007811\times\text{VOR700}-0.00001967\times\text{QAd700}\\&-0.00243\times\text{T500}-0.00002900\times\text{TAd850}\\&-0.00001702\times\text{QAd850}+0.00029288\times\text{NE850}\\&-0.00024344\times\text{RH700}+0.00001071\times\text{PCWT}\\&+0.00228\times\text{SSI}-0.00000694\times\text{CBAS}\\&-0.00037976\times\text{VV500}-0.00047052\times\text{VV850}\\&+0.00027782\times\text{KIDX}+0.00044022\times\text{ET87}\ .\end{aligned}$$

The distribution of the estimated value $\hat{p}$ is also right-skewed. In a similar way, we decide that the threshold T is 0.0202 subject to the same condition: $1-S_1 \approx 0.1$. Table 11 and Table 12 show the results of the model training and model validation respectively.

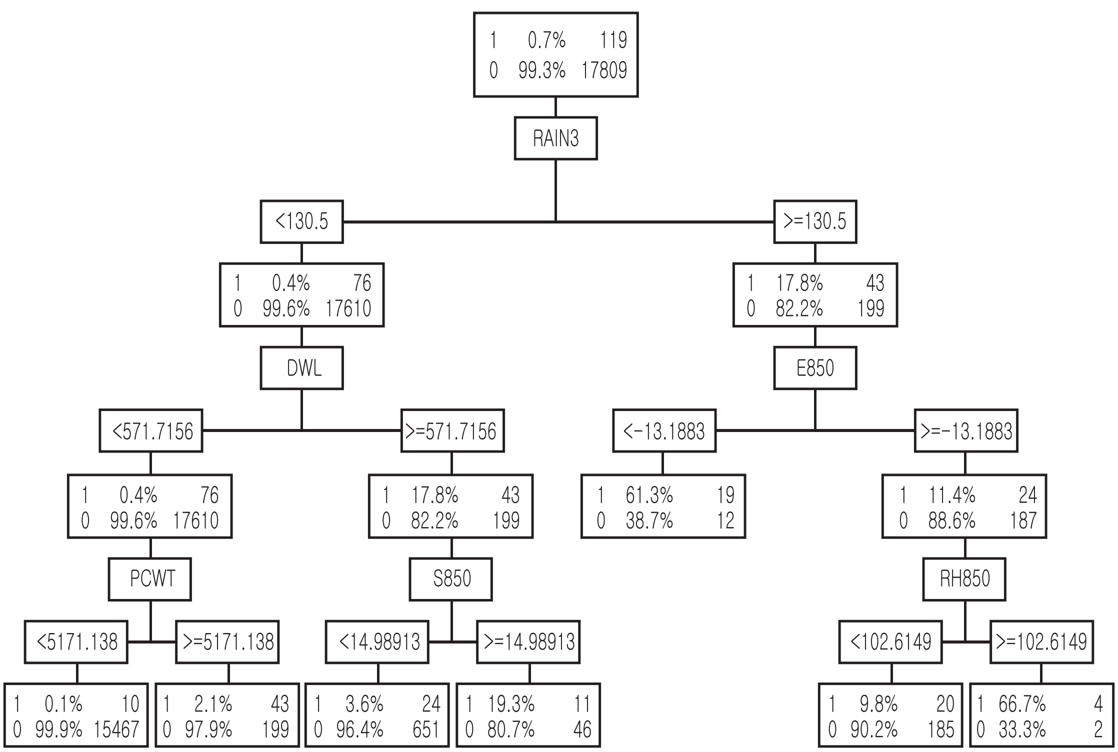

Fig. 6. The final CART model.

Table 11. 2×2 table for the multiple linear regression model training.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(17809 cases)	16027(89.99%)	1782(10.01%)
Heavy rain(119 cases)	19(15.97%)	100(84.03%)

Table 12. 2×2 table for the multiple linear regression model validation.

Observation	Forecast	
	No heavy rain(0)	Heavy rain(1)
No heavy rain(5775 cases)	4978(86.20%)	797(13.80%)
Heavy rain(87 cases)	8(9.20%)	79(90.80%)

Table 13. Comparison by Heidke skill score.

Model	Training	Validation
Logistic Regression	0.09149	0.22267
Neurl Network	0.10869	0.14674
CART	0.07309	0.16002
Linear Regression	0.08857	0.14087

4. Discussion

The "no heavy rain" cases (17 809 cases) comprise 99.3% of the total cases (17 928 cases) in Table 4. This means that the correction rate is 99.3% even if a forecaster always says we will not have heavy rain. In this case, the correction rate is not important. Therefore any forecast model should be compared against a reference forecast which is easier to prepare than the model forecast. The Heidke skill score (HSS) is useful to compare forecast models that generate categorical forecasts. The HSS uses the random forecasts as its reference and removes the effect of the random forecasts. The random forecast, as a kind of reference forecast, is a random variable with the same statistical properties as the predictand. The HSS of Table 2 is defined by

$$\mathrm{HSS}=\frac{A+D-R}{A+B+C+D-R}\,, \quad (5)$$

where

$$R=\frac{(A+B)(A+C)+(B+D)(C+D)}{A+B+C+D-R}\,.$$

If you want to know about skill scores in detail, see Von Storch and Zwiers (1999).

The comparison of the four models via the HSS is summarized in Table 13. Although the neural network model has the best results for the training case, the logistic regression model has the best results for the validation case. Based on the 2×2 tables of model training and validation and the HSS table, the logistic regression model is recommended. We also checked another logistic regression model which includes the main effects and interaction effects of the predictors. However its results are not good (its HSS is 0.08801 for the model training, and 0.16664 for the model validation). For the neural network models, the analysis with only the variables used for the logistic regression model (PCWT, Rain3, S500, T500, VOR700) and the analysis with only the variables used for the CART model (Rain3, DWL, PCWT, S850, E850, RH850) are also performed. For the first one, the HSS is 0.10193 for the model training case, and 0.15064 for the model validation case. For the second one, the HSS is 0.09486 for the model training case, and 0.18369 for the model validation case.

As further works, these statistical models will be applied to the local prediction of climate factors like rainfall, heavy rainfall and snowfall in order to find the optimal forecast model.

Acknowledgments. This research was performed for the project, "Development of Techniques for Local Prediction", for the Research and Development on Meteorology and Seismology funded by the Korea Meteorological Administration (KMA) in 2004.

REFERENCES

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees.* Chapman & Hall, New York, 358pp.

Cho, J. Y., and J. T. Choi, 1995: Probability of Precipitation using Statistical Method. KMA/NWPD Technical Report 95-4, Korea Meteorological Administration, 59pp.

DeNeef, P., and D. L. Kent, 1993: Using treatment-tradeoff preferences to select diagnostic strategies: Linking the ROC curve to threshold analysis. *Medical Decision Making*, **13**, 126–132.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Haykin, S., 1999: *Neural Networks, 2nd Edition.* Prentice-Hall, New Jersey,

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, **8**, 301–349, 842pp.

ISDR, and WMO, 2004: *Water and Disasters: Be Informed and Be Prepared.* WMO publication No. 971. Geneva, Switzerland, World Meteorological Organization, 34pp.

Kim, H. J., H. J. Baek, W. T. Kwon, and B. S. Choi, 2001: Long-range forecast of precipitation using an interval arithmetic neural network. *Journal of the Korean Meteological Society*, **37**, 5443-5452.

Kok, K., and S. Kruizinga, 1992: Updating probabilistic MOS equations. *Proc. 12th Conf. on Probability and Statistics in Atmospheric Sciences*, Versailles, France, Amer. Meteor. Soc., 62–65.

Lemcke, C., and S. Kruizinga, 1988: Model output statistics forecasts (three years of operational experience in the Netherlands). *Mon. Wea. Rev.*, **116**, 1077–1090.

Myers, R., D. C. Montgomery, and G. G. Vining, 2002: *Generalized Linear Models.*, John Wiley & Sons, New York, 342pp.

Raubertas, R. F., L.E. Rodewald, S. G. Humiston, and P.G. Szilagyi, 1994: ROC curves for classification trees. *Medical Decision Making*, **14**, 169–174.

Ross, G. H., 1989: Model output statistics using an updatable scheme, *Proceedings of the 11th Conferences on Probability and Statistics in Atmospheric Sciences*, Monterey, California, American Meteorologist Society, 93–97.

Sohn, K. T., and J. H. Kim, 2003: Statistical prediction of precipitation during warm season in Seoul area. *Journal of the Korean Data Analysis Society*, **5**(1), 113–126. (in Korean)

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, Cambridge, 484pp.