ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 26, NO. 2, 2009, 327–332

Ternary Forecast of Heavy Snowfall in the Honam Area, Korea

Keon Tae SOHN*[1], Jeong Hyeong LEE[2], and Young Seuk CHO[1]

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

[2] *Division of Management Information Science, Dong-A University, Busan* 604-714*, Korea*

(Received 10 January 2008; revised 11 September 2008)

ABSTRACT

The objective of this study is to improve the statistical modeling for the ternary forecast of heavy snowfall in the Honam area in Korea. The ternary forecast of heavy snowfall consists of one of three values, 0 for less than 50 mm, 1 for an advisory (50–150 mm), and 2 for a warning (more than 150 mm). For our study, the observed daily snow amounts and the numerical model outputs for 45 synoptic factors at 17 stations in the Honam area during 5 years (2001 to 2005) are used as observations and potential predictors respectively. For statistical modeling and validation, the data set is divided into training data and validation data by cluster analysis. A multi-grade logistic regression model and neural networks are separately applied to generate the probabilities of three categories based on the model output statistic (MOS) method. Two models are estimated by the training data and tested by the validation data. Based on the estimated probabilities, three thresholds are chosen to generate ternary forecasts. The results are summarized in 3×3 contingency tables and the results of the three-grade logistic regression model are compared to those of the neural networks model. According to the model training and model validation results, the estimated three-grade logistic regression model is recommended as a ternary forecast model for heavy snowfall in the Honam area.

Key words: ternary forecast of heavy snow, MOS, multi-grade logistic regression, neural networks, threshold

Citation: Sohn, K. T., J. H. Lee, and Y. S. Cho, 2009: Ternary forecast of heavy snowfall in the Honam area, Korea. *Adv. Atmos. Sci.*, **26**(2), 327–332, doi: 10.1007/s00376-009-0327-2.

1. Introduction

In South Korea, the property damage caused by heavy snowfall occupied 7.6 percent of the total damage caused by water-related disasters during the 10 years from 1993 to 2002. Heavy snowfall damages agricultural products and establishments, and causes traffic paralysis, etc. According to the geographical characteristics, we have heavy snowfall during every cold season in the Honam area which is located on the southwest side of the Korean Peninsula. Therefore, to reduce the damage, it is important to improve the forecast modeling for heavy snowfall in the Honam area.

There are various types of snowfall forecasts; binary forecasts, multi-categorical forecasts, probabilities of occurrence, probabilities of classified snowfall, and quantitative snowfall forecasts. For example, the binary forecast of heavy snowfall has two-class forecast, "we will have heavy snow tomorrow" or "we will not have heavy snow". The multi-categorical forecast is preferred over the binary forecast because the multi-categorical forecast gives us more detailed information than the binary forecast.

In weather forecast systems, the probabilistic forecast is generally preferred because the probabilities of categories (for instances, the probability of precipitation and the probability of classified precipitation) can easily show the uncertainty of the forecasts as Murphy (1993) commented. Choi and Cho (2002) studied the objective prediction of the probability of precipitation in Korea based on the perfect prognostic method. Sohn and Kim (2003) proposed some statistical models for the probabilities of classified precipitation during the warm season in the Seoul area. However the probabilistic forecasts cannot be used directly if they are estimated too smoothly. In these cases, the categorical forecast is used instead of the probabilistic forecast. Commonly, categorical forecasts can be generated by some thresholds, which are chosen based on the dis-

*Corresponding author: Keon Tae SOHN, ktsohn@pusan.ac.kr

Table 1. Frequency of observations for each station.

	Station																	
Category	140	146	156	165	168	169	170	175	243	244	245	247	248	256	260	261	262	Total
0	475	476	467	480	482	479	480	479	467	464	464	470	465	477	480	480	482	8067 (98.45%)
1	6	5	13	2	0	3	2	3	13	17	13	12	16	5	2	2	0	114 (1.39%)
2	1	1	2	0	0	0	0	0	2	1	5	0	1	0	0	0	0	13 (0.16%)
Total	482	482	482	482	482	482	482	482	482	482	482	482	482	482	482	482	482	8194

tribution of the estimated probabilities. According to Sohn et al. (2005a) for the quantitative rainfall forecast in the Honam area, the estimated probabilities show a tendency to be underestimated. Sohn et al. (2005c) then presented the binary forecast strategy. Similarly, Sohn et al. (2005b) also studied statistical models for the binary forecast of heavy snowfall.

The goal of this study is the development of ternary forecasts of heavy snowfall in the Honam area. In this study, the ternary data has one of the following three values, 0 for (daily new amount of snow cover is less than 50 mm) or 1 for (50–150 mm) or 2 for (more than 150 mm). Two statistical models (a three-grade logistic regression model and a three-grade neural network model) are separately applied to the ternary forecast of heavy snowfall based on the model output statistic (MOS). The MOS, proposed by Glahn and Lowry (1972), is a physical-statistical modeling technique used to find the statistical relationship between the numerical model outputs and the observations. Many authors have considered the MOS to predict the temperature and precipitation (for instances, Lemcke and Kruizinga, 1988; Ross and Studwicke, 1994; Kok and Kruizinga, 1992; Sohn and Kim, 2003; Sohn et al., 2005a,b,c).

In section 2, the predict and and potential predictors are introduced and the model training data and validation data are classified by cluster analysis. In section 3, the forecast modeling strategy for the ternary forecast is presented. In order to estimate the probabilities of the three categories, a multi-grade logistic regression model (Myers et al., 2002) is applied. A neural network model used in Sohn et al. (2005b), which consists of an input layer, one hidden layer, and an output layer, is also considered. In section 4, two models are estimated by the training data and checked by the validation data, separately. Based on the estimated probabilities of the three categories, three thresholds are chosen in order to generate the ternary forecasts. The results are summarized in 3×3 contingency tables and two models are compared. In addition, some concluding remarks are presented in section 5.

2. Data

For our study, the observed daily snow cover and numerical model outputs for 45 synoptic factors at 17 stations in the Honam area, during the cold season (November to March) in 2002 to 2005, are used. Cho and Choi (1995) found that the climatic characteristics of the warm season are different from those of the cold season in Korea, so we only considered the cold season.

In the Korean Meteorological Administration (KMA), the special report for heavy snowfall consists of an advisory (daily new amount of snow cover is more than 50 mm) and a warning (more than 200 mm). We used 150 mm instead of 200 mm because there is only one warning case in the data period. The daily snowfall observations are transformed into ternary values; 0 for less than 50 mm, 1 for 50–150 mm and 2 for more than 150 mm. The frequencies of the three categories (0, 1, 2) are given in Table 1 for each station and Table 2 for each year.

The 45 synoptic factors in Table 3 are used as potential predictors. Sohn et al. (2005a) used these

Table 2. Frequency of observations for each year, training data and validation data.

	Year							
Category	2001	2002	2003	2004	2005	Total	Training	Validation
0	765	1534	1989	2323	1456	8067	4745	3322
1	0	13	41	39	21	114	70	44
2	0	0	10	1	2	13	5	8
Total	765	1547	2040	2363	1479	8194	4820	3374

Table 3. Numerical model outputs as potential predictors.

Symbols	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	North-west wind speed at 850 hPa, 700 hPa and 500 hPa
NE850, NE700, NE500	North-east wind speed at 850 hPa, 700 hPa and 500 hPa
VV850, VV700, VV500	Wind speed at 850 hPa, 700 hPa and 500 hPa
VOR850, VOR00, VOR500	Relative vorticity at 850 hPa, 700 hPa and 500 hPa
QAD850, QAD700	Advection of specific humidity at 850 hPa and 700 hPa
Q84, Q74	Difference of specific humidity at 850 hPa and 700 hPa, at 700 hPa and 700 hPa
TAD850, TAD700	Thermal advection at 850 hPa and 700 hPa
RH850, RH700, RH500	Relative humidity at 850 hPa, 700 hPa and 500 hPa
CCL, DWL, PCWT	Convective condensation level, Depth of wet level, Potential precipitation
CTOP, CBAS, BBX1, BBX2	Level of cloud top, Level of cloud base, Black box index 1, Black box index 2
SSI, KYID, KIDX	Showalt stability index, KY index, K index
LR87, LR85	Lapse rate between 850 hPa and 700 hPa, between 850 hPa and 500 hPa
T850, T700, T500	Temperature at 850 hPa, 700hPa and 500 hPa
ET850, ET700	Equivalent potential temperature at 850 hPa and 700 hPa
ET87	Difference of equivalent potential temperature at 850 hPa and 700 hPa

synoptic factors: the wind direction and speed, relative vorticity, humidity, thermal advection, potential precipitation, and temperatures. All of them can be generated by the numerical model, called RDAPS (Regional Data Assimilation and Prediction System), used by the KMA.

In order to divide the data into the model training data and the validation data, cluster analysis is applied using the new daily snow cover amounts for each station. As a result, data of 10 stations are used for the model training data, and those of the remaining 7 stations are used for the model validation data. Table 2 includes the frequencies of the model training data and the validation data.

3. Forecast strategy for ternary forecasting

3.1 *Forecast models*

The objective of this study is to develop one ternary forecast model to use for heavy snowfall in the Honam area. Ternary forecasting consists of two steps. The first step is to generate the corresponding probabilities of the three categories, and the second is to generate the ternary forecast using the thresholds. In order to generate the corresponding probabilities of the three categories, two statistical models, a three-grade logistic regression model and a three-grade neural network model, are applied and their results are compared. The output of the three-grade models should be a multinomial type vector, which consists of three probabilities, (p_0, p_1, p_2) where p_0 is for Category 0, p_1 is for Category 1, p_2 is for Category 2, and $p_0 + p_1 + p_2 = 1$.

The three-grade logistic regression model for the ordinal and ternary responses is defined by the following two equations:

$$\log\left(\frac{P(Y \leqslant j|\boldsymbol{X}=\boldsymbol{x})}{1-P(Y \leqslant j|\boldsymbol{X}=\boldsymbol{x})}\right) = b_j+\boldsymbol{\beta}'\boldsymbol{x}\ ,\ \ j=0,1\ ,$$

where Y is the observed response (0 or 1 or 2), $\boldsymbol{X}$ is the vector of the predictors, b_j is a constant (an intercept), and $\boldsymbol{\beta}$ is the coefficient vector of the predictors. The significant predictors are selected by the stepwise selection method, and the parameters in the above model are estimated using the training data. The probabilities of the three categories are computed by the following equations.

$$\begin{aligned} p_0 &= P(Y \leqslant 0)\ , \\ p_1 &= P(Y \leqslant 1) - P(Y \leqslant 0)\ , \\ p_2 &= 1 - P(Y \leqslant 1)\ . \end{aligned}$$

The three-grade neural network model, which consists of 45 inputs, one hidden layer, and one output layer, is considered. Similar to Sohn et al. (2005a), the linear basis function is used as a combination function and the logistic function is used as an activation function. However, the final activation function in the output layer is the three-grade logistic function. Optimal weights are estimated via a back propagation algorithm (Haykin, 1999). The number of nodes in

Table 4. Simple forecast strategy using a three-grade logistic regression model (The value in the parenthesis indicates row percentage).

	Forecast			
Observation	0	1	2	Total
0	4735 (99.79%)	10 (0.21%)	0 (0%)	4745
1	56 (80%)	14 (20%)	0 (0%)	70
2	0 (0%)	5 (100%)	0 (0%)	5
Total	4791	29	0	4820

Table 5. Model training results of the three-grade logistic regression: T1=0.02, T2=0.6, T3=0.6.

	Forecast			
Observation	0	1	2	Total
0	4354 (91.76%)	388 (8.18%)	3 (0.06%)	4745
1	6 (8.57%)	59 (84.29%)	5 (7.14%)	70
2	0 (0.0%)	0 (0.0%)	5 (100%)	5
Total	4360	447	13	4820

the hidden layer is determined by the Akaike information criterion (Akaike, 1974). For this study, we used the statistical package, SAS/E-Miner.

3.2 *Thresholds for generating ternary forecasts*

As a simple forecast strategy, it seems reasonable that we choose the category with the maximal value among the three probabilities. In this case of the three-grade logistic regression model, the results of the model training are summarized in the 3×3 contingency table (observation forecast) given in Table 4. Though the exact forecast rate of Category 0 is 99.79%, it has no meaning because Category 1 and Category 2 are much more serious than Category 0. The exact forecast rate for Category 1 is only 20% and there is no forecast rate for Category 2. Therefore we decided to use another strategy with thresholds.

With varying numbers and values of thresholds from 0 to 1 and the mode of inequality, the many 3×3 contingency tables are made and compared. As heuristic results, we decided that three thresholds are needed and the ternary forecasts can be generated by the following algorithm.

IF $(1-P_0) \leqslant$ T1 THEN Forecast = 0;

ELSE IF (T1$\leqslant(1-P_0)<$T2) or ($P_1 \leqslant$T3) THEN Forecast=1;

ELSE Forecast=2;

where T1, T2 and T3 are thresholds.

4. Results

4.1 *Three-grade logistic regression model*

Using the model training data, the two equations of the three-grade logistic regression model are estimated as follows:

$$\begin{aligned} b_0 + \hat{\boldsymbol{\beta}}' \boldsymbol{X} =& 99.3411 + 0.00698 \times \mathrm{CTOP} + \\ & 0.00438 \times \mathrm{DWL} - 0.4744 \times \mathrm{ET700} - \\ & 0.0485 \times \mathrm{NE700} + 0.0238 \times \mathrm{NW500} + \\ & 0.00438 \times \mathrm{DWL} - 0.4744 \times \mathrm{ET700} - \\ & 0.0485 \times \mathrm{NE700} - 0.5065 \times \mathrm{NW850} - \\ & 0.1876 \times \mathrm{S500} - 0.4125 \times \mathrm{S850} - \\ & 0.4583 \times \mathrm{T650} + 0.5889 \times \mathrm{VV850} \end{aligned}$$

and

$$b_1 + \hat{\boldsymbol{\beta}}' \boldsymbol{X} = 96.0712 + 0.00698 \times \mathrm{CTOP} +$$

Table 6. Model validation results of the three-grade logistic regression: T1=0.02, T2=0.6, and T3=0.6.

	Forecast			
Observation	0	1	2	Total
0	3000 (90.31%)	314 (9.45%)	8 (0.24%)	3322
1	3 (6.82%)	39 (88.64%)	2 (4.55%)	44
2	0 (0.0%)	6 (75.0%)	2 (25.0%)	8
Total	3003	359	12	3374

$$0.00438 \times \mathrm{DWL} - 0.4744 \times \mathrm{ET700} -$$
$$0.0485 \times \mathrm{NE700} + 0.0238 \times \mathrm{NW500} +$$
$$0.00438 \times \mathrm{DWL} - 0.4744 \times \mathrm{ET700} -$$
$$0.0485 \times \mathrm{NE700} - 0.5065 \times \mathrm{NW850} -$$
$$0.1876 \times \mathrm{S500} - 0.4125 \times \mathrm{S850} -$$
$$0.4583 \times \mathrm{T650} + 0.5889 \times \mathrm{VV850}.$$

As mentioned in section 3.2, the optimal thresholds are found by monitoring the 3×3 contingency tables for values of the three thresholds from 0 to 1. According to heuristic comparisons, (T1=0.02, T2=0.6, T3=0.6) are chosen as three thresholds for generating the ternary forecast based on the estimated three-grade logistic regression model. That is, the forecasts have ternary values: 0 if $(1-P_0)$ is less than 0.02, 1 else if $(1-P_0)$ is less than 0.6 or P_1 is greater than or equal to 0.6, and 2 otherwise.

The model training results, using these thresholds, are summarized in Table 5. The exact forecast rate is 91.66% totally, 91.76% for Category 0, 84.29% for Category 1, and 100% for Category 2. These results are much better than the simple forecast case in Table 4. The model validation results of this case are summarized in Table 6. The exact forecast rate is 90.13% totally, 90.31% for Category 0, 88.64% for Category 1, and 25% for Category 2. These results are similar to those of the model training case.

4.2 *Three-grade neural networks*

The number of nodes in the hidden layer is determined by minimizing the model identification statistic called the Akaike Information Criterion (AIC). Table 7 shows that the case of the 4-node hidden layer is optimal.

Similar to the logistic regression case, (T1=0.001, T2=0.09, T3=0.9) are chosen as three thresholds for generating the ternary forecast based on the estimated three-grade neural networks model.

The model training results of this case are summarized in Table 8. The exact forecast rate is 93.65% totally, 93.72% for Category 0, 90% for Category 1, and 80% for Category 2. The model validation results of this case are summarized in Table 9. The exact forecast rate is 91.82% totally, 92.38% for Category 0, 63.64% for Category 1, and 12.5% for Category 2. These results are much worse than to those of the model training case for Category 1 and Category 2.

Table 7. AIC value for each node.

Number of node	2	3	4	5	6	7
AIC	627	580	573	673	760	796

4.3 *Comparison of the two models*

In order to check the predictability of the multi-categorical forecast model, Burrows (1991) considered the modified Heidke skill score (mHSS), which is weighted on the distance between the observed category and the forecasted category. See von Storch and Zwiers (1999) or Burrows (1991) for more about the mHSS in detail. The values of the mHSS' are 0.25733 for the three-grade logistic regression and 0.33153 for the neural networks in the model training cases. The values of the mHSS' are 0.21847 for the three-grade logistic regression and 0.34911 for the neural networks in the model validation cases. With the mHSS criterion, the three-grade neural network model is better than the three-grade logistic regression model.

However, as mentioned in section 3.2, Category 1 and Category 2 are much more serious than Category 0. Table 5, Table 6, Table 8, and Table 9 show that the three-grade logistic regression model can give a more accurate forecast of Category 1 and Category 2. Therefore, we recommend the three-grade logistic regression model.

5. Concluding Remarks

Nowadays, people want to know more detailed information on the categorical forecast. Many binary forecasts have changed to multi-categorical forecasts in the KMA. The objective of this study is the development of ternary forecast of heavy snowfall in the Honam area. A three-grade logistic regression model and a three-grade neural networks model are separately applied to the ternary forecast of heavy snowfall based on the model output statistic (MOS). To generate ternary forecasts, thresholds are needed. With heuristic experiments (monitoring the 3×3 contingency tables and varying the values of the three thresholds from 0 to 1), three thresholds are determined for each model. The results of the model training and the model validation cases are summarized in 3×3 contingency tables for each model.

Even though the neural networks models usually give much better results than the regression models in the model training cases because of their nonlinearity, this situation is often broken in the model validation cases. Most forecasters like the logistic regression model even though the results of neural networks are slightly better than that of the logistic regression model, because the physical interpretations on the synoptic weights in the neural networks are much more difficult than the logistic regression model.

With a synthetic view based on the above 3×3 contingency tables, we recommend the three-grade logistic regression model be used for ternary forecast in the

Table 8. Model training results of the three-grade neural networks: T1=0.001, T2=0.09, and T3=0.9.

	Forecast			
Observation	0	1	2	Total
0	4447 (93.72%)	298 (6.28%)	0 (0.00%)	4745
1	0 (0.0%)	63 (90.0%)	7 (10.0%)	70
2	0 (0.0%)	1 (20.0%)	4 (80.0%)	5
Total	4447	362	11	4820

Table 9. Model validation results of the three-grade neural networks: T1=0.001, T2=0.09, and T3=0.9.

	Forecast			
Observation	0	1	2	Total
0	3069 (92.38%)	246 (7.41%)	7 (0.21%)	3322
1	8 (18.18%)	28 (63.64%)	8 (18.18%)	44
2	1 (12.50%)	6 (75.00%)	1 (12.50%)	8
Total	3078	280	16	3374

Honam area. This forecast strategy of using thresholds can be applied to any ternary forecast in weather forecasting systems.

Acknowledgements. This research was performed for the project "Development of technique for Local Prediction", one of the research and development projects on meteorology and seismology funded by the Korea Meteorological Administration (KMA), 2005.

REFERENCES

Akaike, H., 1974: A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Burrows, W. R., 1991: Objective guidance for 0–24 hour and 24–48 hour mesoscale forecasts of lake-effect snow using CART. *Wea. Forecasting*, **6**, 357–378.

Cho, J. Y., and J. T. Choi, 1995: Probability of precipitation using statistical method. KMA/NWPD Technical Report 95–4, Korea Meteorological Administration, 59.

Choi, J. T., and J. Y. Cho, 2002: The objective prediction of probability of precipitation (PoP) based on PPM. *Journal of the Korean Meteorological Society*, **38**, 119–127. (in Korean)

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 ed, Prentice-Hall, New Jersey, 842pp.

Kok, K., and S. Kruizinga, 1992: Updating probabilistic MOS equations. *Proc. 12th Conferences on Probability and Statistics in Atmospheric Sciences*, Amer. Meteoro. Soc., Totonto, Canada, 62–65.

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

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

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

Ross, G. H., and C. C. Studwicke, 1994: Logistic regression using a Kalman filter within an updateable MOS forecasting system. *Proc. 13th Conferences on Probability and Statistics in Atmospheric Sciences*, Amer. Meteoro. Soc., San Francisco, USA, 204–209.

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**, 113–126. (in Korean with English abstracts)

Sohn, K. T., J. H. Lee, and C. S. Ryu, 2005a: Statistical models for quantitative precipitation forecast in the Honam area. *Journal of the Korean Data Analysis Society*, **7**, 507–521. (in Korean with English abstract)

Sohn, K. T., J. H. Lee, and C. S. Ryu, 2005b: Statistical prediction of heavy snow in the Honam area, Korea. *Proc. Fifth METRI-IAP Joint Workshop*, Yanji, China.

Sohn, K. T., J. H. Lee, S. H. Lee, and C. S. Ryu, 2005c: Statistical models for prediction of heavy rain in the honam area. *Korean Journal of the Atmospheric Sciences*, **41**, 897–907. (in Korean with English abstracts)

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