The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 369-378

Binary Forecast of Heavy Snow Using Statistical Models[1)]

Keon Tae Sohn[2)]

Abstract

This Study focuses on the binary forecast of occurrence of heavy snow in Honam area based on the MOS(model output statistic) method. For our study daily amount of snow cover at 17 stations during the cold season (November to March) in 2001 to 2005 and Corresponding 45 RDAPS outputs are used. Logistic regression model and neural networks are applied to predict the probability of occurrence of Heavy snow. Based on the distribution of estimated probabilities, optimal thresholds are determined via true skill score. According to the results of comparison the logistic regression model is recommended.

Keywords : Binary forecast; Heavy snow; MOS; logistic regression; Neural networks

1. 서론

기상예보에서 특정 기상상태의 발생유무를 예보하는 경우가 많다. '발생함'과 '발생하지 않음'으로 구분되는 이 범주 기상예보에 대하여 일반적으로 '내일 비올 확률이 70%이다' 형태인 확률예보를 선호하고 있다. Murphy (1993)에 의하면 확률적 예측을 하는 이유는 예측을 각 범주가 일어날 확률로 나타냄으로써 예측의 불확실성을 한 눈에 나타낼 수 있다는 이점을 가지고 있기 때문이다. 발생확률을 추정하기 위하여 다양한 통계 모형이 사용되고 있으나, 추정된 발생비율이 과대 평활화된 경우에는 모형에서 추정된 발생확률을 그대로 예보치로 사용할 수 없다. 이 경우 차선책으로 문턱치(threshold)를 정하여 발생확률이 문턱치보다 크면 '발생함'으로, 그렇지 않은 경우는 '발생하지 않음'으로 예보하게 되어 이 범주 예보(binary forecast)가 이루어진다.

이 범주 예보는 주로 강수예보에서 많이 발생한다. 강수예보는 강수발생확률예보, 강수발생예보, 범주별 강수확률예보, 강수범주예보, 강수량예보 등으로 구분된다. 최준태와 조주영 (2002)은 강수발생확률예보에 대한 연구를 수행하였다. Sohn 등 (2005)은 집중호우 발생예보 모형에 대하여 연구하였으며, 서울지역 범주별 강수확률예보에 대한 연구로는 손건태와 김재환 (2003)이 있다. 손건태 등 (2005b)의 호남지역 강수량예보에 대한 연구와 손건태 등 (2005a)의 상태종속모형을 이용한 세 시간 강수량예보가

1) 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
2) Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea. E-mail : ktsohn@pusan.ac.kr

있다.

최근 들어 급격한 기상변화로 인한 피해가 전 세계적으로 급증하고 있다. 우리나라에서 10년간(1993년~2002년) 기상과 관련된 재산피해 중 7.6%가 대설에 의한 피해로 보고되고 있다. 특히 호남지역은 지리적 특성으로 강설량이 많아 폭설로 인한 농작물 피해, 시설 파괴, 교통두절 및 막대한 경제적 재산피해를 가져왔으며 해마다 피해의 정도가 커지고 있다.

본 연구에서는 호남지방에 대한 대설발생예보를 위한 이 범주 통계 모형 개발을 목적으로 하였다. 슈퍼컴퓨터에서 생산되는 수치모형 출력자료를 이용한 객관적 해석법으로 역학-통계 모형화 기법 중 하나인 MOS(model output statistics)를 적용하여 수치모형 예보자료와 관측치의 관계를 통계모형으로 추정하여 발생확률 예측모형을 개발하고자 하였다. MOS는 수치모형 예측치들과 관측치 사이의 통계적 함수관계를 규명하여 예측치를 생산하는 기법으로 수치모형에 의한 예측치 생성 작업의 후처리 모형으로 사용할 수 있어 많은 학자들(Glahn과 Lowry, 1972; Lemcke와 Kruizinga, 1988; Ross와 Studwicke, 1994; Kok과 Kruizinga, 1992; 손건태와 김재환, 2003; Sohn 등, 2003; 손건태, 2004; Sohn 등, 2005)에 의하여 기상예측모형 개발에 적용되고 있으나, 수치모형 구조식이 변경되는 경우 예측모형을 재추정해야 하는 단점도 있다.

2절에서는 연구에 사용된 자료에 대하여 설명하였으며, 3절에서는 예측모형 개발 전략과 문턱치 선정 및 예보생성 전략을 기술하였으며, 대설발생예보를 위한 예측모형으로 로지스틱 회귀모형과 신경회로망을 적용하고 비교하였다. 각 모형에 의한 발생확률의 분포를 고려하여 예측성 평가측도(skill score)를 최대로 하는 문턱치를 결정하였으며 결과를 비교하였다. 4절에서 결과를 요약하였다.

2. 자료

예측모형 개발을 위하여 2002년 1월 1일부터 2005년 3월 31일까지 한후기(11월~3월)에 호남지역 17개 지점(군산, 전주, 광주, 목포, 여수, 흑산도, 완도, 진도, 부안, 임실, 정읍, 남원, 장수, 순천, 장흥, 해남, 고흥)에서 각각 482일씩 총 8,194일에 해당하는 일신적설량(daily new snow cover) 관측치와 수치모형 RDAPS(regional data assimilation and prediction system) 예측값들을 모형개발에 사용하였다.

기상청 기상특보기준에 따라 일신적설량 50mm 이상을 '대설발생'으로 50mm 미만인 경우를 '발생하지 않음'으로 정의하였다. 각 지점별 자료 수는 <표 1>, 연도별 자료 수는 <표 2>와 같다. RDAPS 예측값 중 최준태와 조주영 (2002)에서 사용된 45종류의 잠재적 예측인자를 <표 3>에 요약하였다. 결측이 있는 경우는 분석에서 제외하였다.

본 연구에서는 지점마다 별로로 모형식을 추정하는 것이 아니라 17개 지점 자료를 사용하여 호남지역 전체에 대하여 하나의 모형식으로 대설발생유무에 대한 예보모형을 개발하는 것이 연구방향이다. 훈련자료와 검증자료를 구분하기 위하여 군집분석을 사용하였다. 단순임의추출로 훈련자료와 검증자료를 구분하는 방법은 집중호우가 발생된 경우가 1.55% 정도 밖에 되지 않으므로 적용할 수 없으며, 일반적으로 시계열의

앞부분을 모형훈련자료로 뒷부분을 모형검증자료로 사용하기도 하지만 <표 2>에서 보듯이 기간이 짧고 연도별 차이가 많아 적용하기가 문제가 있다고 판단하였다. 훈련자료와 검증자료 모두에서 어느 정도 대설발생 경우가 포함되어야 하며, <표 3>의 잠재적 예측인자가 포함된 전체 변수를 사용하지 않고 일적설량 만을 이용하여 17개 지점에 대하여 군집분석을 수행하고 훈련자료와 검증자료를 구분하기로 하였다. 아주 이질적인 집단으로 훈련자료와 검증자료를 구분하면 모형 훈련자료가 호남지역 전체에 대한 대표성을 상실하고 수치모형 결과 외에 지역적 특성이 모수추정에 영향을 주기 때문에, 군집분석 결과인 <그림 1>의 덴드로그램을 통하여 훈련자료와 검증자료가 각각 전체에 대한 대표성을 어느 정도 지닐 수 있도록 구분하였다. 훈련자료로 {광주, 흑산도, 완도, 진도, 부안, 순천, 해남, 고흥, 남원, 임실}의 10개 지점자료(총 4,820일)를, 검증자료로 {군산, 전주, 장수, 정읍, 목포, 여수, 장흥}의 7개 지점자료(총 3,374일)를 사용하였다.

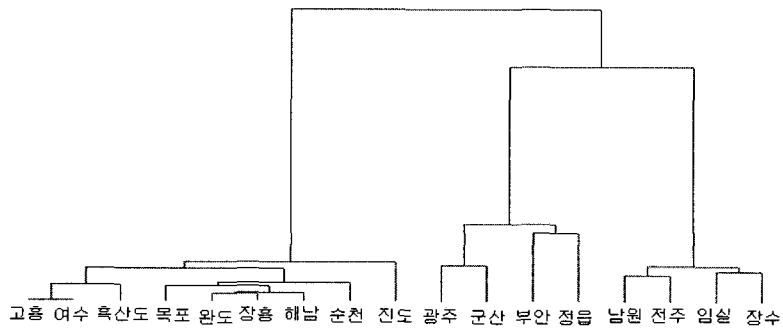
<표 1> 각 지점에 대한 관측치 빈도표

지점	140	146	156	165	168	169	170	175	243
0	475	476	467	480	482	479	480	479	467
1	7	6	15	2	0	3	2	3	15
계	482	482	482	482	482	482	482	482	482

지점	244	245	247	248	256	260	261	262	Total
0	464	464	470	465	477	480	480	482	8067
1	18	18	12	17	5	2	2	0	127
계	482	482	482	482	482	482	482	482	8194

<표 2> 년도별 관측치 빈도표

년도	2001	2002	2003	2004	2005	Total
0	765	1534	1989	2323	1456	8067
1	0	13	51	40	23	127
계	765	1547	2040	2363	1479	8194



<그림 1> 덴드로그램

<표 3> 잠재적 예측인자들

기호	예측인자
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,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 of specific humidity at 850 hPa and 700 hPa
Q74	Difference of 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
BBX	Black box index
SSI	Showalter stability index
KYID	KY index
KIDX	K index
LR87	Lapse rate between 850 hPa and 700 hPa
LR8	Lapse rate between 850 hPa and 500 hP
T850,T700,T500	Temperature at 850 hPa, 700hPa and 500 hPa
ET850,ET700	Equivalent potential temperature at 850 hPa and 700 hPa
ET8	Difference of equivalent potential temperature at 850 hPa and 700 hPa

3. 예측모형 개발

3.1 예보 생성 전략

호남지역 전체에 대하여 하나의 모형으로 예측모형을 추정하기로 하였다. 예측모형으로 로지스틱 회귀모형과 신경회로망을 별도로 적용하였으며 모형훈련자료를 사용하여 각 모형을 추정하였다.

로지스틱 회귀모형에서 단계별 변수선택을 적용하였다. 고려된 신경회로망 구조는 (1) 각 예측인자(RDAPS 예측치)의 표준화 자료를 입력하는 입력층, (2) 선형기저함수를 정보전달함수로 하고 로지스틱 함수를 정보활성화 함수로 하는 1개 은닉층, (3) 로지스틱 함수를 정보활성화 함수로 하는 출력층(발생확률생산)으로 이루어진다. 역전파 알고리즘에서 평균오차를 최소화하는 최적모형 선택기준을 적용하였으며, 모형식별통계량 Akaike information criterion(AIC)을 최소로 하는 노드 수를 선정하였다.

Sohn 등 (2005)의 집중호우발생예보 연구에서 보듯이 강설 경우에도 모형에서 추정된 발생확률 분포가 0쪽으로 치우쳐있어 예측된 발생확률을 직접 예보에 사용하기 어려우므로 문턱치를 고려하여 대설발생유무를 결정하였다.

이 범주 관측과 예보의 결과는 <표 4>의 2×2 분할표로 요약된다. 이 범주 예보모형의 예측성 평가측도(skill score)는 2×2 분할표에 기초하여 계산되며, 선택된 평가측도가 최대가 되는 문턱치를 결정하여 예보 생산에 사용한다. 이 범주 예보모형에 대한 예측성 평가측도는 <표 4>의 2×2 분할표를 사용하여 계산된다. 이 범주 및 다범주 예측성 평가 측도들에 대한 소개와 연구는 손건태와 한정임 (2004)과 Hans와 Francis (1999)를 참고하기 바란다.

문턱치 결정 후 모형훈련자료와 모형검증자료에 대하여 추정된 모형과 결정된 문턱치를 사용하여 이 범주 예보를 생성하고 관측과 예보로 이루어진 두 모형의 2×2 분할표를 비교하여 예보모형을 결정하는 과정으로 연구가 수행되었다. 모형추정 작업은 SAS E-Miner로 수행되었으며 문턱치에 변화에 따른 예보 생성 및 2×2 분할표 작성과 비교는 SAS/MACRO 프로그래밍으로 하였다.

<표 4> 이 범주 예보에 대한 2×2 분할표

관측	예보		계
	발생하지 않음(0)	발생함(1)	
발생하지 않음(0)	D(correct negative)	B(false alarm)	B+D
발생함(1)	C(miss)	A(hit)	A+C
계	C+D	A+B	A+B+C+D

3.2 모형훈련 및 모형검증 결과

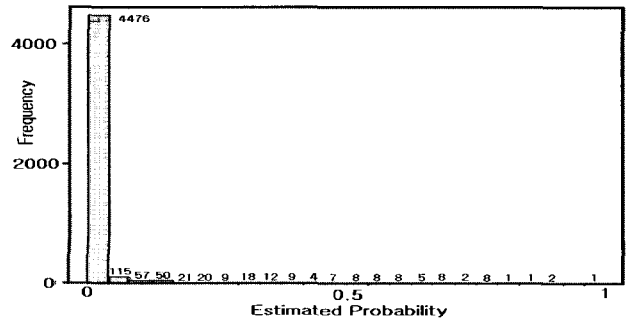
3.2.1 로지스틱 회귀모형 적용 결과

단계별 변수선택법을 적용하여 추정된 모형식은 다음과 같다.

$$\hat{p} = \frac{1}{1+\exp(-h)}$$

여기서 h 는 다음 식으로 계산된다.

$$\begin{aligned} h = \; & 90.7512 + 0.00732 \times CCL + 0.00513 \times CTOP + 0.00742 \times DWL \\ & - 0.4077 \times ET700 + 0.0276 \times NW500 - 0.1812 \times S500 \\ & - 0.4243 \times T850 + 0.1019 \times VV850 \end{aligned}$$



<그림 2> 로지스틱 회귀모형 적합 시 추정된 발생확률의 분포

발생확률의 분포가 <그림 2>의 형태를 가지며, 추정된 로지스틱 회귀모형으로 예측된 발생확률이 0.5보다 큰 경우 대설발생으로 예보한다면 <표 5>의 결과가 나온다. 즉, 문턱치를 0.5로 선정한 결과이다. 실제 대설발생인 경우에 대한 예보 정확도가 더 중요하므로 18.67%인 예보를 제공할 수 없다. 그렇다고 발생확률을 그대로 예보하기에도 확률값이 작아 문제가 있다. 따라서 알맞은 문턱치를 결정해야할 필요가 있다.

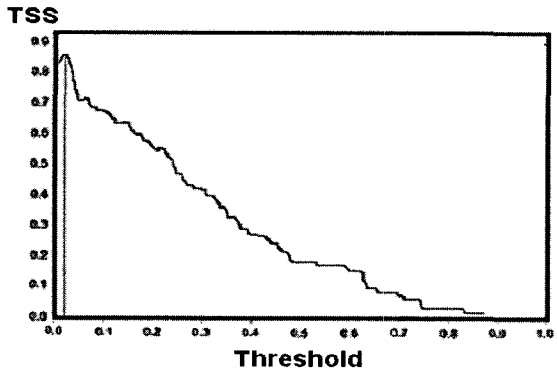
<표 5> 2×2 분할표 (로지스틱 회귀모형, 문턱치 = 0.5)

관측	예보		계
	0	1	
0	4733(99.75%)	12(0.25%)	4745
1	61(81.33%)	14(18.67%)	75
계	4794	26	4820

Sohn (2006)에 의하면 이 범주 예보모형에 대한 예측성 평가측도는 발생확률 분포와 발생비율에 따라 다르게 선정되어야 하므로, 제안된 예측성 평가측도 선정 가이던스에 따라 true skill statistic을 선택하여 문턱치를 결정하였다. True skill statistic은 <표 4>의 2×2 분할표를 사용하여 다음의 식으로 계산된다.

$$TSS = \frac{A}{A+C} - \frac{B}{B+D}$$

문턱치를 0부터 1까지 변화시키며 예보를 생성한 후, 2×2 분할표를 사용하여 계산된 TSS가 <그림 3>에 그려져 있으며 TSS가 최대가 되는 문턱치는 0.02이다. 이때의 예보결과를 모형훈련과 모형검증에 대하여 각각 2×2 분할표로 요약하여 <표 6>에 정리하였다.



<그림 3> 로지스틱 회귀모형 적합 시 각 문턱치에 대한 True skill score 그림

<표 6> 모형훈련과 검증 결과 요약표 (로지스틱 회귀모형, 문턱치=0.02)

		모형 훈련 경우			모형검증 경우		
		예보			예보		
		0	1	계	0	1	계
관측	0	4324 (91.13%)	421 (8.87%)	4745	2982 (89.77%)	340 (10.23%)	3322
	1	4 (5.33%)	71 (94.67%)	75	2 (3.85%)	50 (96.15%)	52
	계	4328	492	4820	2984	390	3374
TSS		0.8580			0.8592		

3.2.2 신경회로망 적용 결과

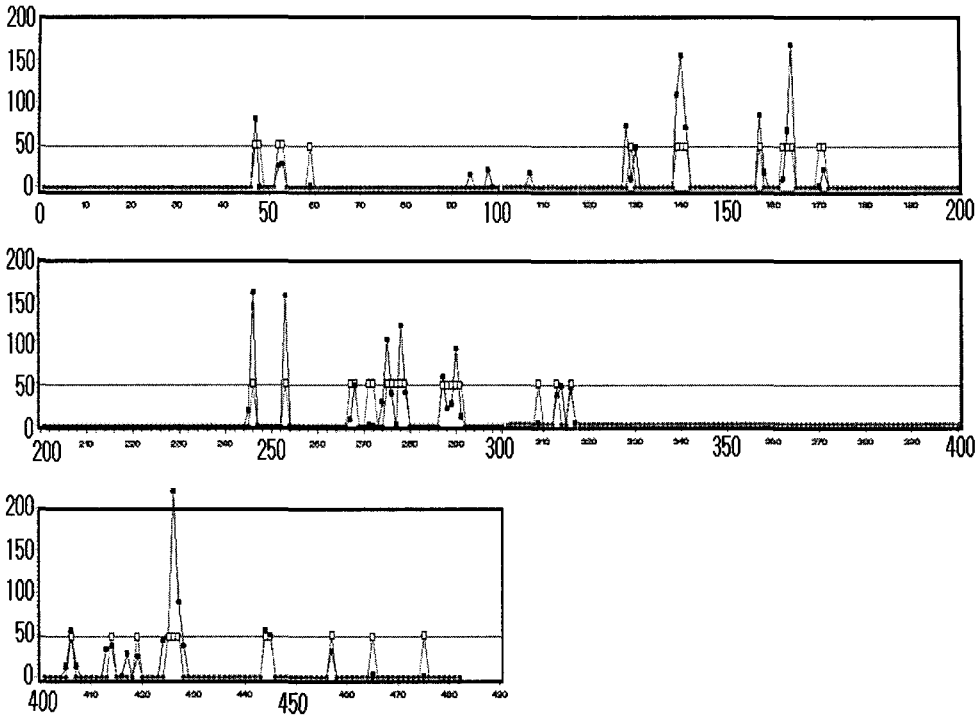
AIC를 최소로 하는 은닉층의 노드 수는 2로 선정하였다. TSS가 최대가 되는 문턱치는 0.01이며, 예보결과를 모형훈련과 모형검증에 대하여 각각 2×2 분할표로 요약하여 <표 7>에 정리하였다.

<표 7> 모형훈련과 검증 결과 요약표 (신경회로망, 문턱치=0.01)

		Training case			Validation case		
		Forecast			Forecast		
		0	1	total	0	1	total
Obs.	0	4484 (94.50%)	261 (5.50%)	4745	3098 (93.26%)	224 (6.74%)	3322
	1	1 (1.33%)	74 (98.67%)	75	4 (7.69%)	48 (92.31%)	52
	total	4485	335	4820	3102	272	3374
TSS		0.9317			0.8557		

3.2.3 모형의 비교

모형훈련에서 2×2 분할표의 각 셀의 비율과 TSS값에 기초하여 살펴보면 신경회로망을 적합한 경우가 로지스틱 회귀모형을 적합한 경우보다 우수한 예보 결과를 보이고 있다. 모형 검증 결과에서는 대설발생 시에는 로지스틱 회귀모형이 더 정확한 예보를 제공하고 있으나 발생하지 않은 경우에는 예측 오류율이 10.23%가 된다. 따라서 예보모형으로 신경회로망의 사용을 제안한다. <그림 4>는 신경회로망을 적용한 예보에서 정읍지역에 대한 모형검증자료 관측치(일 신적설량, ■로 표기)와 이 범주예보(대설발생예보시 참조선 50mm 상에 □로 표기)의 시계열 그림을 작성한 것이다. 참조선 이상이 대설발생 경우이므로 강설이 전혀 없는 경우에는 예보도 '발생없음'으로 하고 있으며, 시계열 그림에서 볼 때 예측모형과 문턱치에 의한 예보전략이 어느 정도 정확도를 유지하고 있음을 알 수 있다.



<그림 4> 정읍지역에 대한 일 신적설량 관측치(■)와 이 범주 예보치(□)에 대한 시계열그림: (신경회로망, 모형검증)

4. 결론

본 연구는 호남지역 이 범주 대설예보를 위하여 로지스틱 회귀모형과 신경회로망을 적용하여 예보모형을 개발한 결과이다. 군집분석을 통하여 훈련자료와 검증자료를 구분하였으며, 두 모형에 의하여 추정된 발생확률의 분포를 바탕으로 문턱치를 사용하여 이 범주 예보를 생산하였다. 두 모형의 예보 결과를 각각 2×2 분할표로 요약한 결과, 로지스틱 회귀모형을 적용한 경우는 예보정확도가 모형훈련에서 91.18%(대설발생시 94.67%), 모형검증에서 89.86%(대설발생시 96.15%)이며, 신경회로망을 적용한 경우는 예보정확도가 모형훈련에서 94.56%(대설발생시 98.67%), 모형검증에서 93.24%(대설발생시 92.31%)로 나타났다.

전체적으로 볼 때 호남지역 이 범주 대설예보를 위한 예측모형으로 신경회로망의 사용과 문턱치를 0.01로 하는 예보 생성을 제안한다. 그러나 추정된 신경회로망에서 입력변수와 발생확률 간의 관련성에 대한 물리적 해석이 어렵기 때문에 예보전문가들은 설명의 폭이 넓은 로지스틱 회귀모형을 선호하므로 예보모형의 선택은 예보담당자의 판단에 맡긴다. 본 논문에서 적용된 모형화 전략은 이 범주 예보가 이루어지는 다른 기상인자들과 다른 지역에도 적용이 가능하다.

참고문헌

[1] 손건태 (2004). 3시간 기상예보를 위한 수치모델 예측치의 통계적 수정. 「Journal of the Korean Data Analysis Society」, 제6권 2호, 453-464.

[2] 손건태, 김재환 (2003). 서울지역 난후기에 대한 통계적 강수예보 모델 개발. 「Journal of the Korean Data Analysis Society」, 제5권 1호, 113-126.

[3] 손건태, 이정형 (2005a). 3시간 강수량예보를 위한 상태종속모형 개발. 「Journal of the Korean Data Analysis Society」, 제7권 1호, 137-150.

[4] 손건태, 이정형, 류찬수 (2005b). 호남지역 강수량 예측을 위한 통계모형 개발. 「Journal of the Korean Data Analysis Society」, 제7권 2호, 507-521.

[5] 손건태, 한정임 (2004). 순서형 확률예측 모형에 대한 새로운 예측성 평가측도의 제안. 「Journal of the Korean Data Analysis Society」, 제6권 1호, 267-278.

[6] 최준태, 조주영 (2002). PPM을 이용한 객관적 강수확률 예보법. 「한국기상학회지」, 제38권 2호, 119-127.

[7] Glahn, H.R., and Lowry, D.A. (1972). The use of model output statistics (MOS) in Objective weather forecasting. *Journal of Applied Meteorology*, Vol. 11, 1203-1211.

[8] Hans, V.S. and W.Z. Francis (1999). *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge.

[9] Kok, K. and Kruizinga, S. (1992). Updating probabilistic MOS equations. *Proceedings of the 12th Conferences on Probability and Statistics in Atmospheric Sciences*, American Meteorologist Society, 62-65.

[10] Lemcke, C. and Kruizinga, S. (1988). Model output statistics (three years of operational experience in the Netherlands). *Monthly Weather Reviews*, Vol. 116, 1077-1090.

[11] Murphy, A.H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting,* Vol. 8, 281-293.

[12] Sohn, K.T. (2006). Guidance on the Use of Skill scores for Two-class Forecast. *Proceedings of the 6th METRI-IAP Joint Workshop*, Jeju, Korea.

[13] Sohn, K.T., Lee, J.H., Lee, S.H. and Ryu, C.S. (2005). Statistical Prediction of Heavy Rain in South Korea. *Advanced in Atmospheric Sciences*, Vol.

22, 703-710.

[14] Sohn, K.T., Rha, D.K. and Seo, Y.K. (2003). The 3-hour-interval prediction of ground-level temperature in South Korea using Dynamic linear model. *Advanced in Atmospheric Sciences*, Vol. 20, 575-582.

[15] Ross, G.H. and Studwicke, C.C. (1994). Logistic regression using a Kalman filter within an updateable MOS forecasting system. *Proceedings of the 13th Conferences on Probability and Statistics in Atmospheric Sciences*, American Meteorologist Society, 204-209.

[Received May 2006, Accepted June 2006]