Verification of Statistical Forecast Guidance for 1-Day and 2-Day Mesoscale Forecasts of Lake-Effect Snow Off Lake Huron, Georgian Bay

W.R. BURROWS
Meteorological Services Research Branch
Atmospheric Environment Service
4905 Dufferin Street
Downsview, Ontario M3H 5T4, Canada

ABSTRACT

A technique has been developed by Burrows (1990b, 1991) for generating 0-24-hour and 24-48-hour mesoscale forecasts of lake-effect snow in five categories for individual stations and small areas using CART. The method was verified for a period independent from the one in which it was developed. The results showed the forecasts to perform relatively well, considering the difficulty of the five-category mesoscale forecast problem. The best success was achieved with forecasts of snow not exceeding 5 cm. The success of forecasts for snow amounts greater than this for specific sites and small areas can be substantially increased when groups of forecasts for small areas are used.

INTRODUCTION

A system for producing guidance for 0-24-hour and 24-48-hour mesoscale forecasts of lake-effect snow for individual stations and small areas was developed by Burrows (1990b, 1991). The technique uses a recently-developed non-parametric classification procedure known as Classification and Regression Trees (CART) (Breiman et al, 1984) to find decision trees which classify categorical snowfalls with threshold values of predictors in binary decision nodes. The probability of snow amount in cm at each station is forecast in five categories: 1 = 0-trace, 2 = >trace-5, 3 = >5-12.5, 4 = >12.5-22.5, 5 = >22.5. These arbitrary categories were chosen to correspond roughly to the southern Ontario public's perception of the severity of snowfall in inches: 1 = no snow, 2 = light snow (up to 2 inches), 3 = light-moderate snow (>2-5 inches), 4 = moderate snow (>5-9 inches), and 5 = heavy snow (>9 inches). Predictors with which to classify the snow events were designed from meteorological parameters known to be important in lake-effect snow formation. Predictors were calculated with 0-24-hour and 24-48-hour forecast data from the Canadian Meteorological Center's operational spectral numerical weather prediction (NWP) model. The stations for which the technique was originally developed lie to the lee of Lake Huron and Georgian Bay, and are shown in Figure 1.

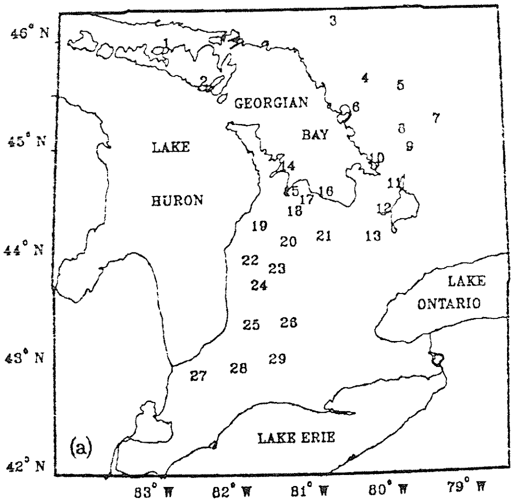


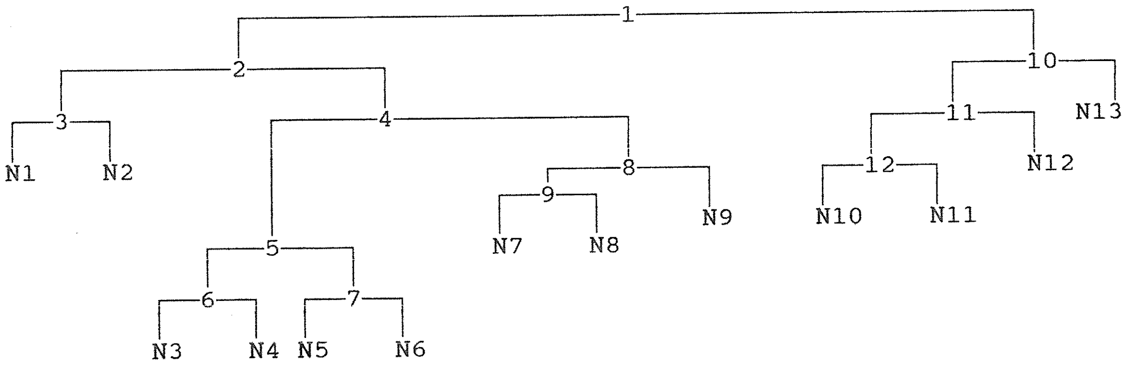
Figure 1: Location of climatological stations included in this study.

An example of the decision tree found for 0-24-hour forecasts at station 19 (Paisley, Ontario) is shown in Table 1. An event enters the tree at node 1 and falls down through the branches to a Terminal Node where it is assigned a final classification category. The latter is taken as the category with the maximum number of events that accumulated in the node when the tree was developed with the "learning data sample". The distribution of those events in the terminal node allows a probability of occurrence to be assigned to the categories in the node. An explanation of CART and its use for this problem has been documented (Burrows 1990b, 1991). The interested reader may refer there for further information.

The CART decision trees were developed with learning data for the November-March winters of 1984-1988. The method has been implemented at the Ontario Weather Center to provide real-time operational forecasts of lake-effect snow. This paper describes verification of the forecasts for a period independent from that used to develop the decision trees.

Table 1: Classification tree diagram for 0-24-hour forecasts at Paisley, Ontario. Nodes in tree paths are numbered. Terminal Nodes at the end of paths in the tree are numbered N1-N13. Data populations in Terminal Nodes appear below.

Tree Structure



Learning Data Populations in Terminal Nodes

Terminal	Assigned	Category				
Node	Category	1	2	3	4	5
N1	2	15	51	11	1	1
N2	3	1	0	4	0	0
N3	3	0	1	5	0	0
N4	5	0	0	0	0	5
N5	4	2	7	5	24	1
N6	3	1	0	4	0	0
N7	2	2	9	2	2	0
N8	3	0	0	4	0	0
N9	3	0	1	14	0	1
N10	2	3	7	1	0	0
N11	1	60	19	4	0	0
N12	2	1	6	0	0	0
N13	2	1	9	0	0	0

VERIFICATION RESULTS

Verification of forecasts was done for days on which lake-effect snow was possible (*LESP days*) for each of 28 stations to the lee of Lake Huron and Georgian Bay for 0-24-hour and 24-48-hour forecasts generated with NWP model predicted data and the CART trees for the independent period November 1988 to March 1989 and November to December 1989. Three verification measures are shown in Figure 2. While these will be defined and discussed below, the reader may wish to refer to Stanski et al (1990) for further discussion.

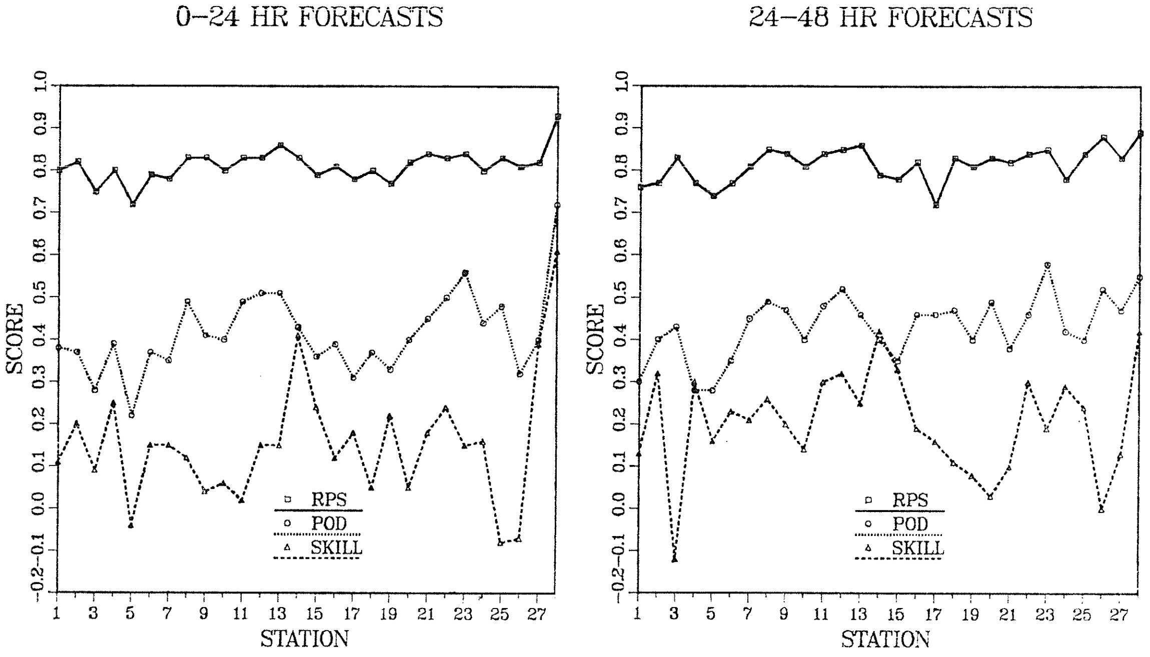


Figure 2: Rank probability score (RPS), probability of detection (POD), and skill score with respect to climate-weighted chance forecasts (SKILL) for CART forecasts for the verification period November 1988-March 1989 and November-December 1989. Because forecasts were not done for station 27 in Fig. 1, stations 28 and 29 there were renumbered 27 and 28, respectively, for this graph.

Rank probability scores (RPS) were calculated with the linear weighting matrix given in Table 2. The method for calculating the RPS is explained in the table heading. The average RPS was .81 for 0-24 forecasts and .82 for 24-48-hour forecasts. The probability of detection (POD) is defined as

$$\text{POD} = \frac{\text{CF}}{\text{O}} \quad , \qquad (1)$$

where CF is the number of events correctly forecast and O is the number of observed events. POD ranges between 0 and 1. The overall POD here averaged .42 for 0-24-hour forecasts and .43 for 24-48-hour forecasts. These verification statistics, while showing the guidance to be far from perfect, demonstrate that it performs reasonably well for both periods, considering the difficulty of the problem and noting that verification was done only for LESP days rather than for the entire data sample. Thus, days on which snow of any kind was unlikely were not included in any of the statistics. These would have been easily forecast correctly. Their inclusion would have substantially increased the scores and hence the "apparent" accuracy of the forecasts.

Table 2: Rank probability score (RPS) weight matrix use to score forecasts against verifying observations. A forecast earns a weight of 1.0 if it agrees with the observation, .75 if it differs by one category from the observation, .50 if it differs by two categories, .25 if it differs by three categories, and 0 if it differs by four categories from the observation. To calculate the RPS, each verified forecast is slotted into its position in the contingency table structure below. Results are summed over the verification period, multiplied by the appropriate weight, then divided by the number of forecasts. The RPS can range between a high of 1.0 and a low of 0.0.

		Observed Category				
		1	2	3	4	5
	1	1.0	.75	.50	.25	0.0
Forecast	2	.75	1.0	.75	.50	.25
Category	3	.50	.75	1.0	.75	.50
	4	.25	.50	.75	1.0	.75
	5	0.0	.25	.50	.75	1.0

The Heidke skill scores in Fig. 2 were calculated with the rank probability scores of CART forecasts and unskilled climate-weighted chance forecasts according to the formula

$$SS = \frac{F - C}{1 - C} \qquad (2)$$

where F is the RPS of the CART forecasts, 1 is the RPS that would be obtained if all the CART forecasts were correct, and C is the RPS of climate-weighted chance forecasts. The latter were determined for every LESP day by assigning a snow category for each station based on a random number between 0 and 1 that was generated for each verification day. Its value was compared to the ranges of cumulative totals of the frequencies of occurrence of the snow categories observed at the station for the verification period. For example, suppose the frequencies of occurrence of the 5 snow categories at a station were .65, .20, .10, .04, and .01. The cumulative frequencies of occurrence would be .65, .85, .95, .99 and 1.00. The random number .43 would be assigned category 1, the random number .97 would be assigned category 4, and so on. A positive score in (2) indicates the CART forecasts to have the greater skill. The average skill scores were .15 and .20, respectively, for the 0-24-hour and 24-48-hour forecasts. This shows the CART forecasts had positive skill with respect to chance forecasts based on climatology.

The above verification statistics were calculated for individual stations without regard to forecasts at nearby stations. If we allow for some leeway in the location of the region of heavy snow in the forecasts relative to the observed, then the forecasts proved to be reasonably good in many instances. Some examples follow. Figures 3 and 4 show the observed snow category and residual (observed category minus forecast category) for three periods of heavy lake effect snow (9 December 1988, 9 February 1989, and 20 December 1989), and for one period of light to moderate lake-effect snow (30 December 1988). A positive residual represents an under-forecast and a negative residual represents an over-forecast. As was previously seen (Burrows 1990b, 1991), CART was not able to classify category 5 snow events at every station, and even category 4 events at some stations. This occurred usually, but not always, in relatively dry locations. Category 4 and 5 snowfalls were reported on all the heavy snow days shown here. While not always forecast exactly where they occurred, they were forecast at nearby stations or somewhere within a small area surrounding the station where they occurred. For example, to the lee of central Georgian Bay for the 9 February 1989 case (Fig. 3), category 5 snow occurred at one station which had a 0-24-hour

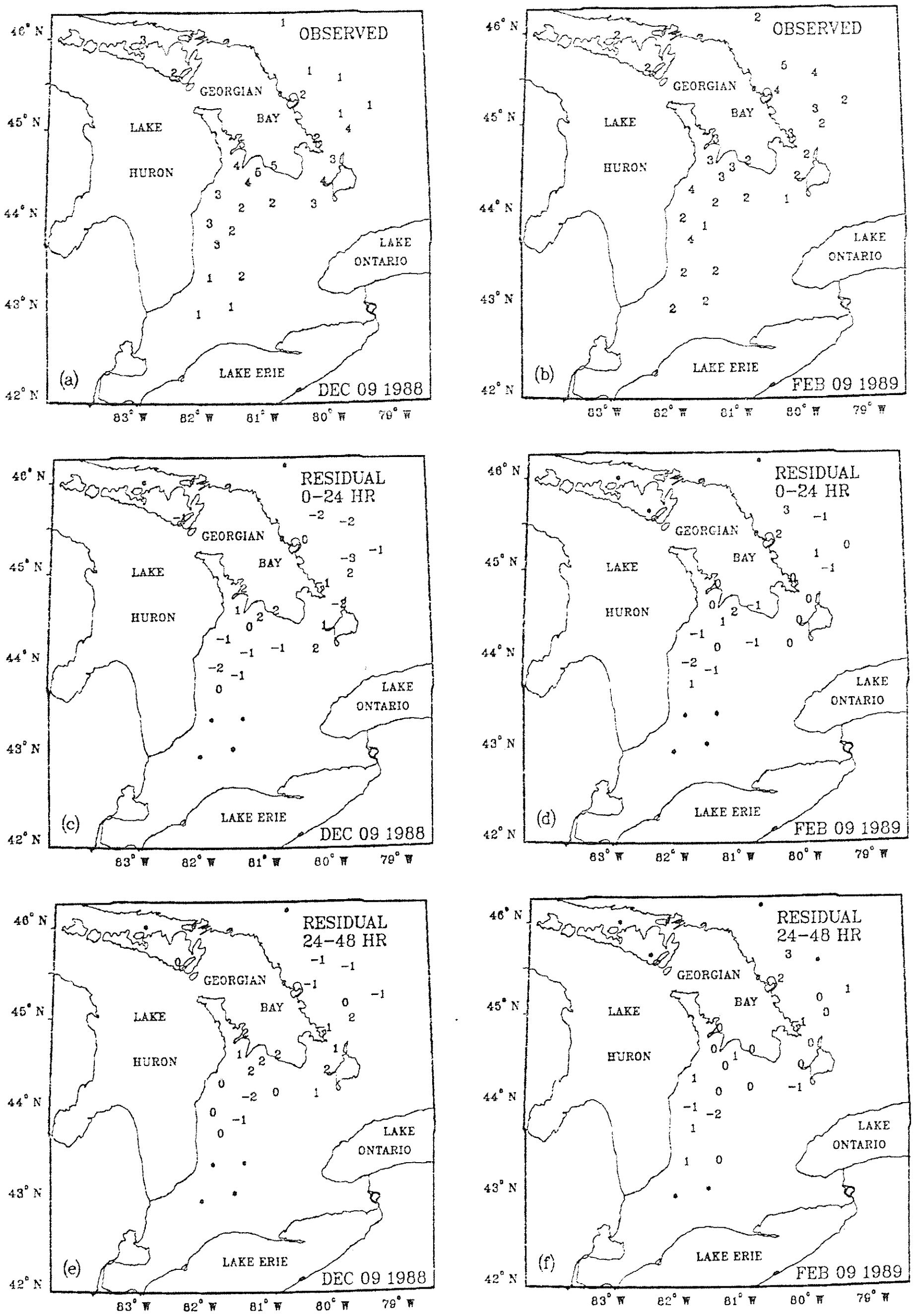


Figure 3: (a) observed snow categories for 9 December 1988; (b) observed snow categories for 9 February 1989; (c) residual (observed snow category minus CART-classified category) for 0-24-hour forecasts valid 9 December 1988 and; (d) same as (c) for 9 February 1989; (e) residual for 24-48-hour forecasts valid 9 December 1988; (f) same as (d) for February 9 1989.

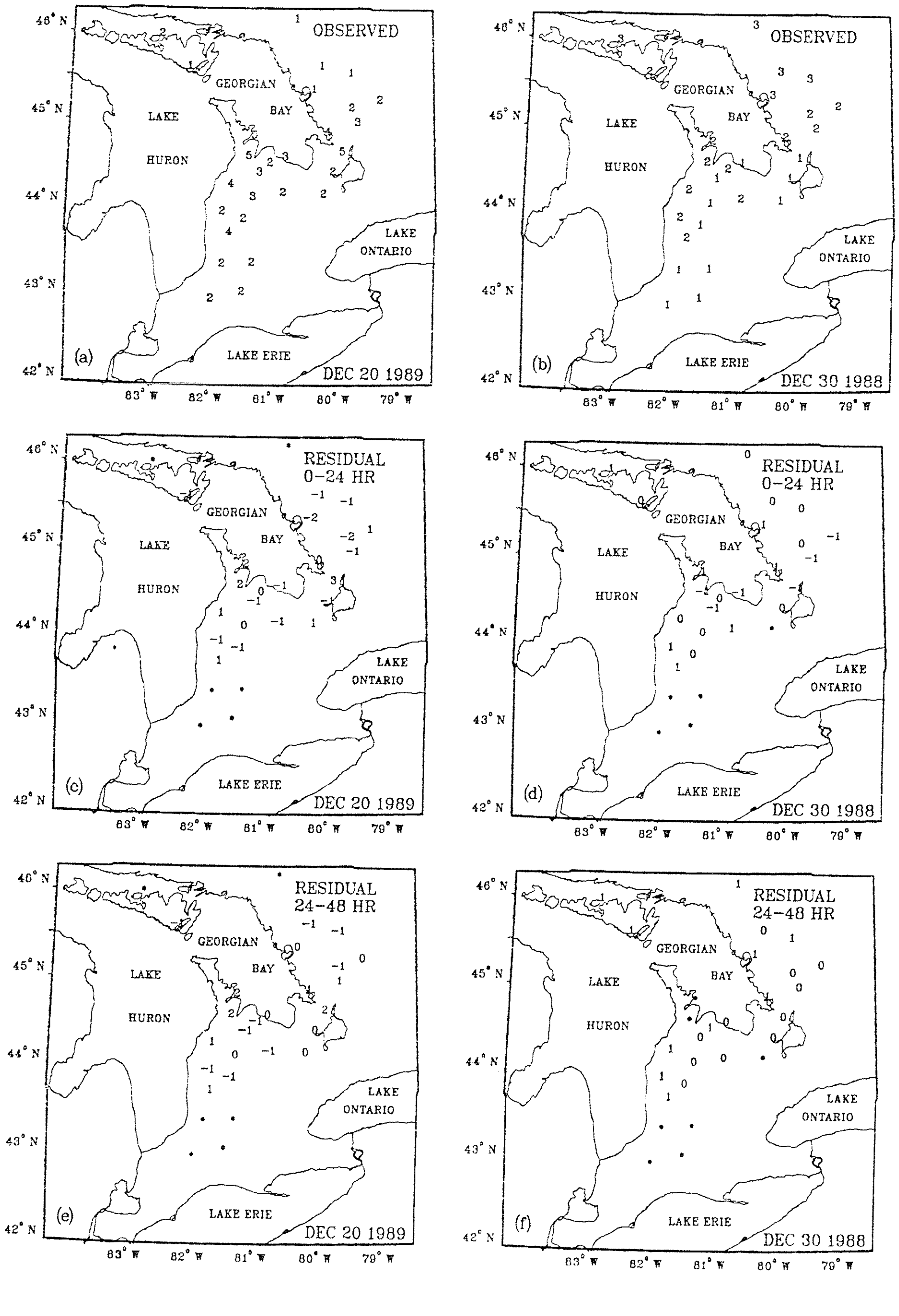


Figure 4: Same as Figure 3, for 20 December 1989 and 30 December 1988.

forecast of category 2 (residual of 3), but a forecast of category 5 *was* made for the station to the southeast, which reported category 4 snow. Similarly, the -3 residual in the 0-24-hour forecast to the lee of Georgian Bay on 9 December 1988 occurred when a forecast of category 4 snow was made for a station which reported category 1 (no snow), but the station immediately to the southeast *did* report category 4 snow. Category 5 snow along the southwestern shore of Georgian Bay on the same day was forecast as category 3 there, but a 0-24-hour forecast of category 5 snow was made to the east, downwind from this region, and also to the south. The case of light to moderate lake-effect snow on 30 December 1988 was generally well forecast everywhere. In all four cases the maximum observed snow category (or within 1 category of it) was forecast somewhere not too far away from where it occurred in both 0-24-hour and 24-48-hour forecasts. There were many other examples that illustrated this.

The critical success index (CSI) gives a measure of the overall usefulness of categorical forecasts. It is defined for a given category or set of categories as

$$\mathrm{CSI} = \frac{\mathrm{CF}}{(\mathrm{F} + \mathrm{O} - \mathrm{CF})} \qquad (3)$$

where CF is the number of correct forecasts, F is the number of forecast events, and O is the number of observed events. The CSI's for the verification period indicated that the greatest success was achieved for forecasts of snow not exceeding 5 cm, i.e. forecasts that category 1 or 2 would occur. (Categories 1 and 2 are the most common snow occurrences at individual stations on LESP days). Respective CSI's for these cases were .66 for 0-24-hour forecasts and .65 for 24-48-hour forecasts, hence about 2/3 of these forecasts were "successful" according to this measure. The CSI's were considerably lower for forecasts of snow categories 3-5, which occur less frequently than categories 1 and 2. However, it is possible to improve the success of forecasts for these categories by using the conglomerate of forecasts over small areas.

As discussed previously, if a certain snow category was forecast at a station and did not occur there, it may have occurred at a nearby station. It seemed likely that using the conglomerate of forecasts for groups of stations covering small areas would prove advantageous, in particular for forecasts of categories 3-5. Stations were grouped by small areas using the official boundaries of Ontario Public Weather Forecast Regions (Small Scale). The station groups are shown in Table 3:

Table 3: Grouping of stations within small areas. Numbers match locations in Figure 1.

Group 1:	1,2,3	Group 4:	14,15,16,17,18,19,20,21
Group 2:	4,5,6,7,8,9	Group 5:	22,23,24,25
Group 3:	10,11,12,13	Group 6:	26,28,29

When the observation at each station was verified against the forecast category value nearest to it in its group, much higher verification scores were found in all categories than was the case when it was verified against the original forecast for that station. This can be seen by noting the consistently higher numbers in the "Nearest Fcst in Grp" column compared to the "Original Fcst" column in Table 4. Thus the correct snow category for a station was often forecast somewhere in a small area surrounding the station, although there is a large decrease in accuracy of forecasts of category 4-5 from the 0-24-hour period to the 24-48-hour period. A strategy is needed for deciding on how to consistently use a group of forecasts. This will depend on the needs of the user. Two suggested strategies follow.

For an *area* forecast, the concern is to forecast snowfall up to a certain depth. We see from the "Grp Max Fcst" column in Table 4 that using the maximum snow category that was forecast within the area's group of stations substantially increases the success of area forecasts in categories 3-5 at almost no loss in RPS or POD. The numbers shown there are averages over the six regions when the maximum snow category that was forecast in each group was selected and verified with the maximum snow category observed in that group.

In forecasts for *individual stations*, a user may require more success in categories 3-5 than the original CART forecasts have. A strategy based on blending the strength of the original CART forecasts in categories 1 and 2 with the strength of the area-grouped forecasts in categories 3-5 is suggested:

1. if the station forecast is less than category 3 and the maximum category forecast in its group is less than category 4, use the *station* forecast.

Table 4: Averages of measures used to verify forecasts with independent data. RPS is the overall rank probability score. POD is the overall probability of detection. "CSI12" is the CSI for forecasts of categories 1-2, where a "hit" occurs when both forecast and verifying observation are in categories 1 or 2. "CSI345" is the critical success index (CSI) for forecasts of categories 3-5, where a "hit" occurs when both forecast and verifying observation are in any of categories 3, 4, or 5. "H3PLUS" is the fraction of observed category 3 events where the forecast was either 3, 4, or 5 (i.e. category 3 was "covered" by a forecast of an equal or higher snowfall). "POD45" is the fraction of events where both the observed and forecast snow categories were either 4 or 5. Columns are organized as follows: "Original Fcst" refers to the forecast made by CART for a station; "Nearest Fcst in Grp" refers to verification of station observation against the forecast category value nearest to it in its group in Table 3; "Blended Fcst" refers to forecasts made for a single station based on consideration of its forecast and the maximum category value that was forecast for all other stations in its group according to the method outlined in Section 2; "Grp Max Fcst" refers to forecasts made for a small area by choosing the maximum category that was forecast within the area's group of stations denoted in Table 3.

Measure	Original Fcst	Nearest Fcst in Grp	Blended Fcst	Grp Max Fcst
0-24-hr				
RPS	.81	.94	.73	.81
POD	.42	.79	.32	.40
CSI12	.66	.87	.44	.53
CSI345	.25	.62	.29	.43
H3PLUS	.38	.68	.69	.62
POD45	.17	.44	.44	.46
24-48-hr				
RPS	.82	.93	.78	.82
POD	.43	.78	.38	.43
CSI12	.65	.87	.44	.58
CSI345	.20	.58	.28	.43
H3PLUS	.30	.61	.65	.57
POD45	.02	.13	.13	.09

2. if the station forecast is equal to category 3 and the maximum category forecast in its group is less than category 5, use the *station* forecast.

3. otherwise, use the *group maximum* forecast.

In Table 4 the blended station forecasts showed a substantial increase of success in categories 3-5 over the original station forecasts, with only a modest decline in the overall average POD and RPS. The decline was mainly due to increased overforecasting of categories 1-3, which should be acceptable for most users.

Table 5: Verification of the observed snow category against the original station forecasts, the blended station forecasts, and the area's group maximum forecasts for station 14 in Group 4 (Wiarton, Ontario airport).

0-24-hour

Observed		Original Fcst					Blended Fcst					Group Max Fcst				
Cat	----->	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Pre-	1	14	13	1	1	0	9	7	0	0	0	0	0	0	0	0
dic-	2	12	26	14	2	0	4	11	6	0	0	13	18	6	0	0
ted	3	2	12	14	5	5	12	20	7	6	0	12	20	7	6	0
	4	0	2	4	2	1	3	13	18	4	6	3	13	18	4	6
Cat	5	0	0	0	0	0	0	2	2	0	0	0	2	2	0	0

24-48-hour

Observed		Original Fcst					Blended Fcst					Group Max Fcst				
Cat	----->	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Pre-	1	12	15	4	0	0	9	11	3	0	0	0	1	0	0	0
dic-	2	9	25	11	5	3	7	9	4	3	1	16	19	7	3	1
ted	3	8	9	15	5	4	12	26	21	5	2	12	26	21	5	2
	4	1	2	2	0	0	2	5	4	2	4	2	5	4	2	4
Cat	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

As an example for a specific station, Table 5 gives the verification results obtained by verifying the station's observed snow category against the original station forecasts, the blended station forecasts, and the area's group maximum forecasts for station 14 in Group 4 (Wiarton, Ontario airport). For this station, CART was not able to separate out category 4 snow from category 5 snow in the classification trees, but category 5 snow occurred during the verification period. We see that the under-forecasting difficulty with the original forecasts in categories 3-5 has been alleviated in the blended and group maximum forecasts, but at the expense of overforecasting in categories 1-3.

CONCLUSIONS

Verification of individual station forecasts with independent data showed the CART-based MOS method outlined here can provide useful, reasonably accurate objective mesoscale guidance for 0-24-hour and 24-48-hour operational forecasts of 24-hour lake-effect snow amount. The best success was achieved with forecasts of snow amount not exceeding 5 cm (categories 1-2). However, the situation often occurred where snow occurrences in categories 3-5 were missed at a station but the forecast for another station nearby would have been much better if it had applied to the first station. It was found that the overall success of forecasts in these categories could be substantially increased if groups of forecasts for small areas were used. Two strategies for using grouped forecasts were discussed. For public forecasts the concern is to forecast snowfall up to a certain depth. Using the maximum category that was forecast among the stations within the area of interest gave a substantial increase in the success of forecasts in categories 3-5, although at the expense of accuracy of forecasts in categories 1-2. In forecasts for individual stations, a user may require more success in categories 3-5 than the original CART forecasts have. A strategy based on blending the strength of the original CART forecasts in categories 1 and 2 with the strength of the area-grouped forecasts in categories 3-5 was suggested and was shown to improve the success of forecasts in these categories.

REFERENCES

Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C.J., 1984: *Classification and Regression Trees*. Wadsworth & Brooks/Cole, Monterrey, 358pp.

Burrows, W.R., 1990a: Tuned perfect prognosis forecasts of mesoscale snowfall for southern Ontario. *J. Geophys. Res.* 95, No. D3: 2127-2141.

Burrows, W.R., 1990b: Objective guidance for 1- and 2- day mesoscale forecasts of lake-effect snow. *Proc. of the 47th Eastern Snow Conference;* Bangor, ME; June 7-8, 1990: 121-134.

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

Stanski, H.R., Wilson, L.W., and Burrows, W.R., 1990: Survey of common verification methods in meteorology. *World Weather Watch Technical Report No. 8,* WMO/TD No. 358. World Meteorological Organization, Geneva, Switzerland, 114pp. (Not available from American Meteorological Society).