Grants ATM-8407142 and 8600543.

An Experiment in Mesoscale Weather Forecasting in the Michigan Area

DENNIS G. BAKER

Department of Atmospheric & Oceanic Science, University of Michigan, Ann Arbor, MI 48109

(Manuscript received 23 July 1986, in final form 25 September 1986)

ABSTRACT

During an experiment in mesoscale weather forecasting in the Michigan area, consensus improved over NWS guidance in maximum/minimum temperature and probability of precipitation forecasts out to 24 hours. Forecasts were generally best in the vicinity of the forecast site. Climatology, persistence, NWS guidance and consensus forecast errors in maximum/minimum temperature and precipitation probability forecasts were divided into synoptic and mesoscale contributions. The errors of the climatology and persistence forecasts resulted substantially more from the synoptic scale than the mesoscale. NWS guidance and consensus forecasts improved over climatology on the synoptic scale with much less or no improvement on the mesoscale. For temperature, consensus showed improvement over guidance only on the synoptic scale. For precipitation, no relationship to scale was evident. In terms of remaining errors, errors were distributed approximately equally between the synoptic scale and mesoscale for temperature. For precipitation, the errors were significantly greater from the mesoscale than the synoptic scale. The implications of these results for zone forecasts and future improvements in mesoscale forecasting accuracy are discussed.

1. Introduction

Of the various forecasts issued by the National Weather Service, probably the one of widest dissemination to the public sector is the zone forecast. Each National Weather Service Forecast Office (WSFO) issues zone forecasts, with a total of over 600 forecasts issued nationwide several times each day. Maps of the zones for the United States are given, for instance, in National Weather Service (1985). As an example, the zones for the lower peninsula of Michigan are shown in Fig. 1. Zones are presumably chosen so that meteorological conditions within each zone are sufficiently homogeneous for the forecast to apply uniformly within the zone. There is a wide variation in the area of zones, with smaller zones in the eastern United States and larger zones in the West.

A WSFO forecaster has to consider the expected weather in his/her area of forecast responsibility (usually a state). States in the Michigan area have an average maximum dimension of 600 km. A typical dimension for a zone in this region is around 100 km. Fujita (1981) proposed that the upper limit of the mesoscale be around 400 km. Thus, the forecaster at a WSFO in the Michigan area has to consider weather features extending from somewhat larger than the mesoscale down well into the mesoscale.

Gedzelman (1981) stated that mesoscale errors impose a limit on weather forecast accuracy. Sanders (1973), Bosart (1975, 1983), Ramage (1976) and Fritsch and Kreitzberg (1978) have suggested that most of the remaining errors in short-range forecasting result from the inability to predict the weather on the mesoscale. On the other hand, Baker (1982) found about equal contributions from the synoptic scale and mesoscale in 24-h MOS (Glahn and Lowry, 1972) maximum/minimum temperature forecast errors for the period 1975–76. This paper describes an experiment designed to examine mesoscale forecasting in the projection period out to 24 hours and the relative contribution of the mesoscale and synoptic scale to forecasting errors.

2. The experiment

Since not all zones in the Michigan area have stations available for verification, the Michigan zones could not be used directly. Instead, major observing stations were used both to insure quality observations and because NWS MOS guidance was available for these stations. The ten stations shown in Fig. 1 were selected.

From 1979 through 1983, students taking synoptic meteorology laboratories at the University of Michigan made forecasts for these ten stations. Since the results for each contest were generally similar, only results for 1980, the contest with the largest number of participants, are given here. For each station, forecasts of the 0000–1200 GMT minimum temperature, 1200–2400 GMT maximum temperature, and the probability of precipitation (PoP) for the 0000–1200 and 1200–2400 GMT periods were made five days a week. Forecasts had to be submitted before 1700 LT. Forecasters had available information from Service A, NAFAX and, late in the term, a GOESFAX circuit. In addition, selected meteorological information was obtained from Weather Services International's interactive data ac-

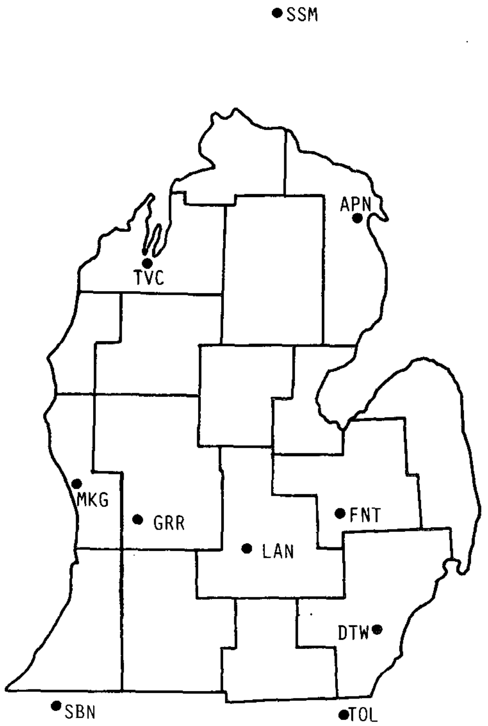


FIG. 1. Forecast zones for lower Michigan with stations for forecast competition. The stations are SSM—Sault Ste. Marie; TVC—Traverse City; APN—Alpena; MKG—Muskegon; GRR—Grand Rapids; LAN—Lansing; FNT—Flint; SBN—South Bend; DTW—Detroit (Metropolitan Airport); and TOL—Toledo.

quisition system. Altogether, 19 forecasters participated, with 8 making 60% or more of the 63 possible forecasts. Six of the eight were senior-level meteorology undergraduates or graduate students. Only two participants (one being the author) would be considered experienced forecasters. Eleven sophomore-level meteorology majors from the introductory synoptic meteorology laboratory made forecasts twice a week. The contest ran from 8 September through 5 December 1980.

Guidance for maximum/minimum temperatures at this time consisted of 24-h local midnight-to-midnight temperature forecasts. Thus, the time periods for guidance and verification were not the same. This fact gave the forecasters an advantage in situations where differences between the weather expected for the guidance and verification periods would be expected. The 3-h MOS temperature forecasts were available to the forecasters but were not used in the verification. Usually, guidance for the 1200 GMT run was used, but when it was not available 0000 GMT output was used.

For each forecast, a consensus forecast was calculated by averaging the forecasts for each particular station and variable. Only the results for consensus are presented here. Overall, consensus did better than all of the individual forecasters. Climatology was obtained from National Climatic Center (1973) for maximum/minimum temperatures, and from Jorgensen (1967) for probabilities of precipitation. A persistence forecast was entered using the observations for the day of the forecast.

3. Verification and presentation of results

For temperature, an rms error is used for verification; namely,

$$T=\frac{1}{N}\left[\sum_{i=1}^{N}(T_i-T_i^{\mathrm{obs}})^2\right]^{1/2}$$

where T_i and T_i^{obs} are the predicted and observed temperatures, respectively, and N is the number of forecasts. For precipitation probability, the Probability Score, B, (one-half of the score defined by Brier, 1950) is used; namely,

$$B=\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)^2$$

where P_i and O_i are the predicted and observed (zero or one) probability of precipitation, respectively, and N is the number of forecasts.

Figures 2 to 5 give the results of the forecasting competition by station and variable. Persistence is used as a control. Although climatology did somewhat better and is used by most researchers, the author feels persistence is a better control in this context because it indicates day-to-day variability. A stable weather regime with large departures from normal would result in large errors for climatology, even though forecasting could be relatively easy. However, persistence errors would be low in such situations. Also, for temperature, no climatology for 12-h periods was available. The left analyses show errors for persistence. The middle analyses give the improvement of guidance over persistence as defined by the skill score (using persistence). For temperature, the skill score would be

$$S_g^{\mathrm{temp}}=\frac{T_p-T_g}{T_p}$$

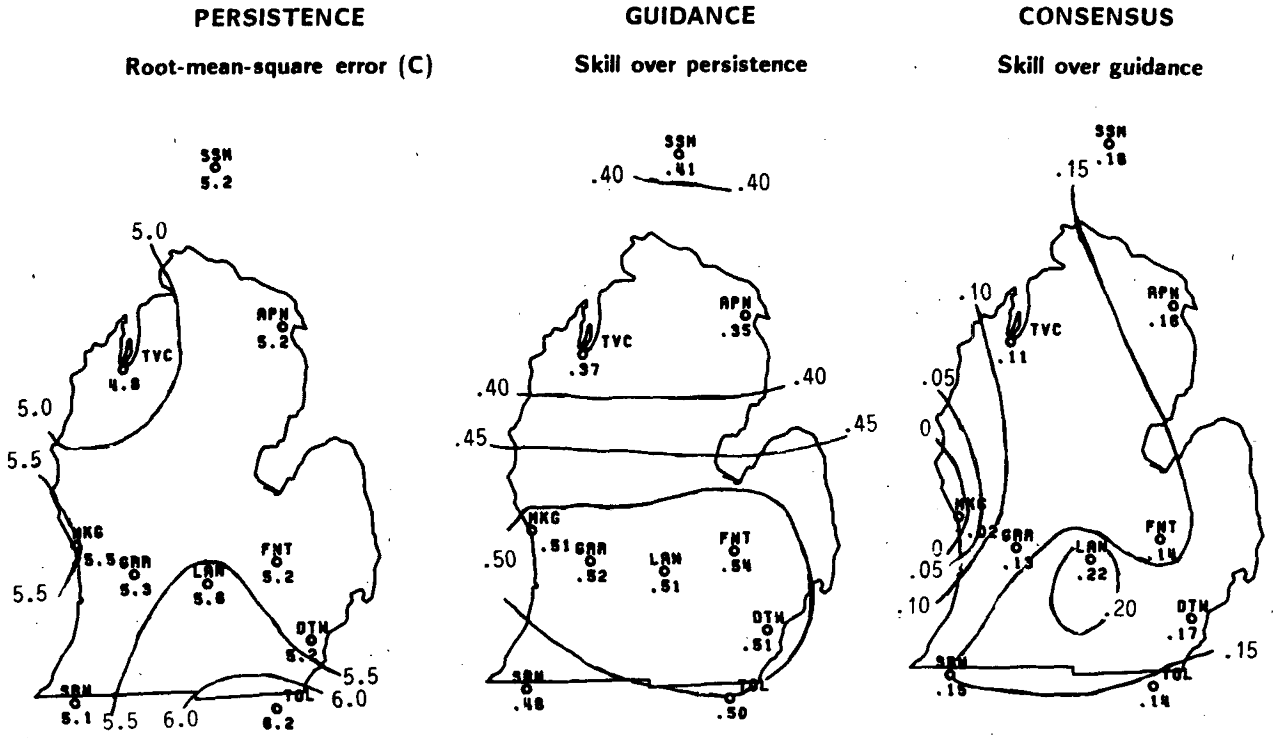
where T_p and T_g are the rms errors for persistence and guidance. For precipitation probability, the skill score would be

$$S_g^{\mathrm{PoP}}=\frac{B_p-B_g}{B_p}$$

where B_p and B_g are the Probability Scores for persistence and guidance. The skill score is negative if guidance does worse than persistence. It would be 1 for perfect forecasts.

The right analyses are the improvement of consensus over guidance defined like the skill score. For temperature, the score would be

$$I_c^{\mathrm{temp}}=\frac{T_g-T_c}{T_g}$$

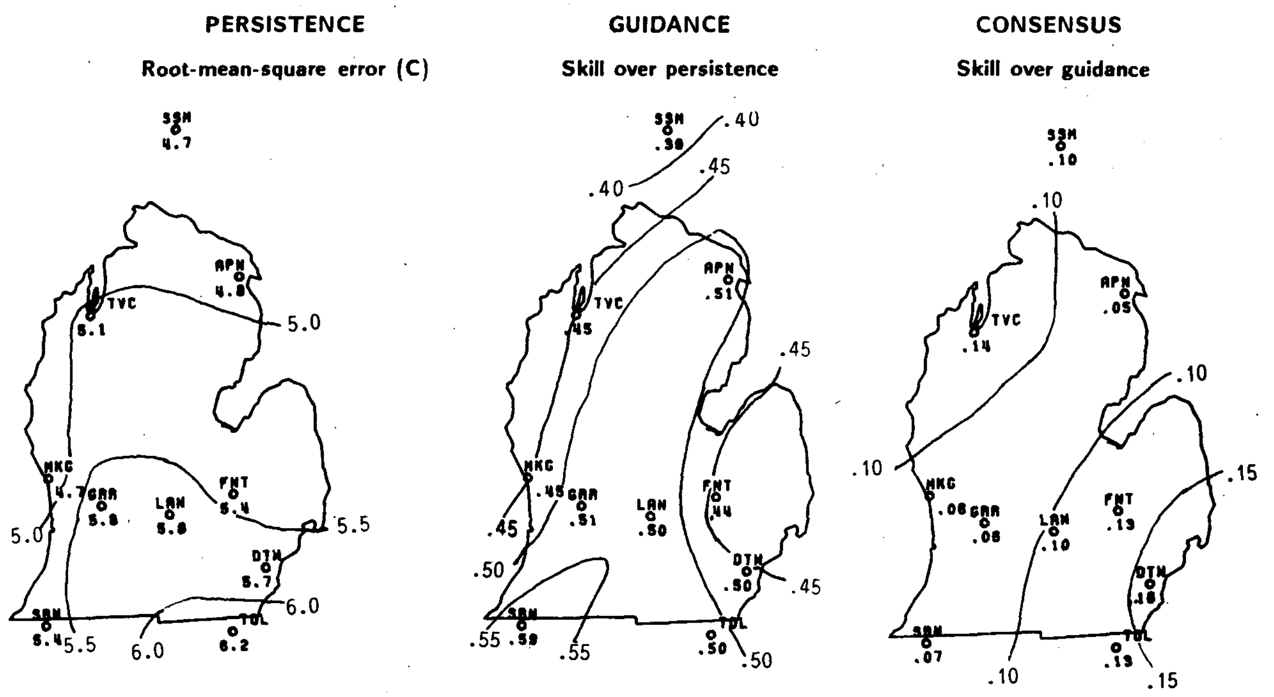


PERSISTENCE
Root-mean-square error (C)

GUIDANCE
Skill over persistence

CONSENSUS
Skill over guidance

Fig. 2. Results for fall 1980 competition for forecasting minimum temperature.



PERSISTENCE
Root-mean-square error (C)

GUIDANCE
Skill over persistence

CONSENSUS
Skill over guidance

Fig. 3. As in Fig. 3 for forecasting maximum temperature.

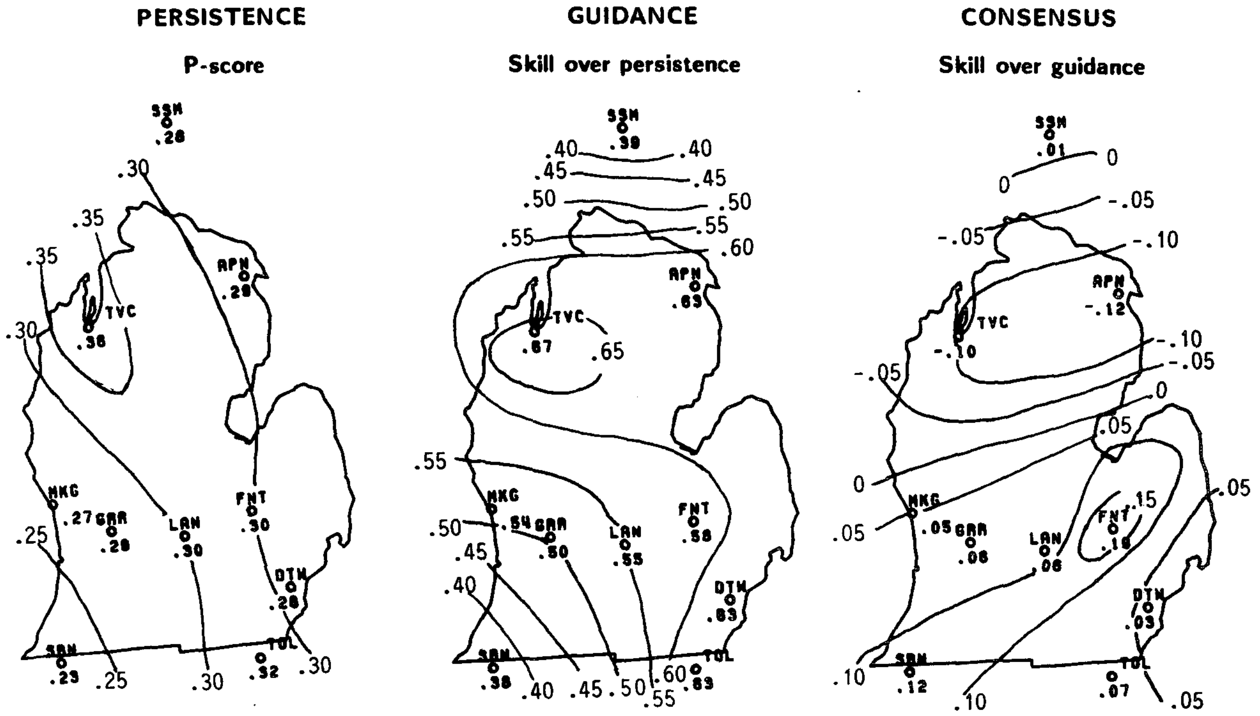


FIG. 4. Results for fall 1980 competition for forecasting 0000–1200 GMT probability of precipitation.

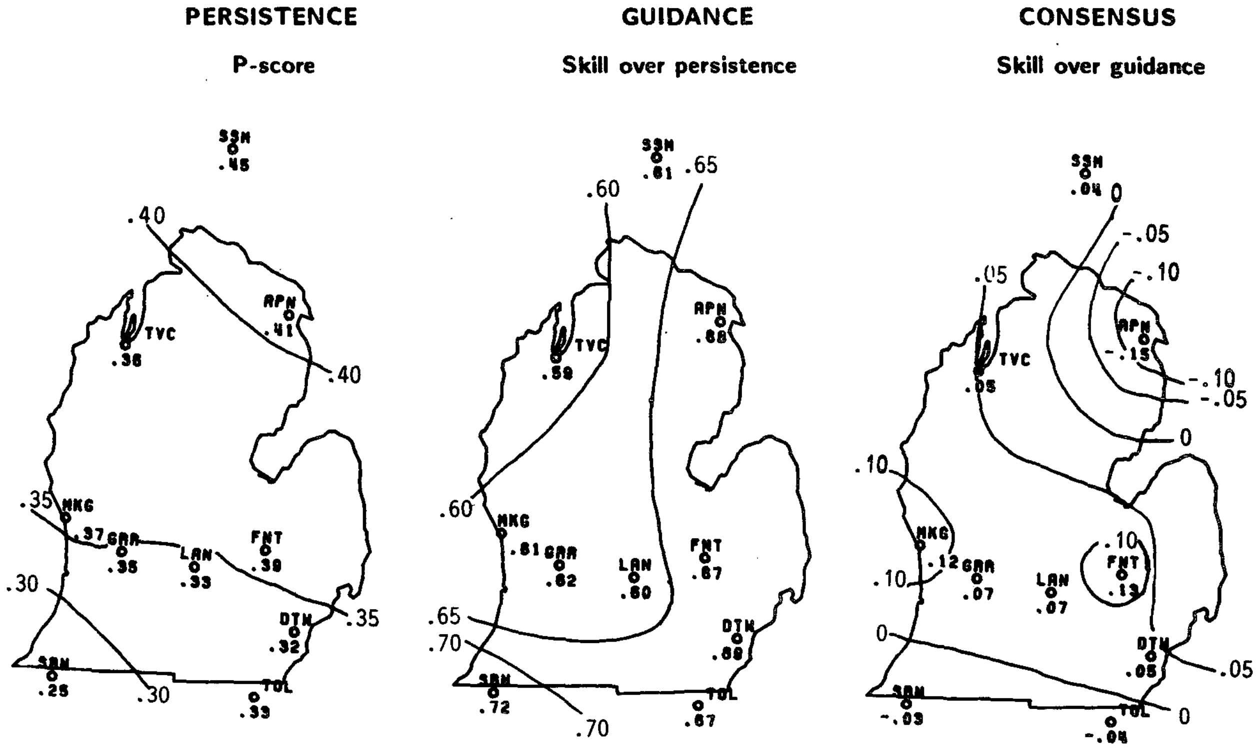


FIG. 5. As in Fig. 4 but for 1200–2400 GMT.

where T_c is the rms error for consensus. For precipitation the score is given by

$$I_c^{\mathrm{PoP}} = \frac{B_g - B_c}{B_g}$$

where B_c is the Probability Score for consensus.

Baker (1982) showed that in an rms verification scheme, the error, E, can be divided into two terms: 1) the average error resulting from the stations collectively, and 2) the deviation error resulting from the variation of each station from the daily network average

$$E^2 = \frac{1}{N}\sum_{j=1}^{N}(\bar{P}_j - \bar{V}_j)^2 + \frac{1}{M \times N}\sum_{j=1}^{N}\sum_{i=1}^{M}(P'_{ij} - V'_{ij})^2$$

where

$$V'_{ij} = V_{ij} - \bar{V}_j \quad P'_{ij} = P_{ij} - \bar{P}_j$$

$$\bar{P}_j = \frac{1}{M}\sum_{i=1}^{M} P_{ij} \quad \bar{V}_j = \frac{1}{M}\sum_{i=1}^{M} V_{ij};$$

P_{ij} is a forecast, V_{ij} is its corresponding verification, and M is the number of stations. For temperature, $E^2 = T^2$, while for precipitation, $E^2 = B$. For temperature, the square root of each term is presented for ease of interpretation. For precipitation, the Probability Score is already squared, so the value of each term is presented as is.

These terms can be interpreted as indicators of spatial scale. The first term comes from factors which affect all the stations. With a maximum distance between two stations of 550 km, this term represents synoptic-scale phenomena. The second term comes from weather events that cause stations to deviate from the area average. In most cases these features would be mesoscale (or less) in extent. Trends across the region (such as large-scale climatological latitudinal variations) appear as mesoscale contributions.

Table 1 gives the results of the experiment overall using the above breakdown. The average rms or Probability Score error is given in the first column, the synoptic-scale contribution in the second column, and the mesoscale contribution in the third column.

4. Results

a. *Maximum/minimum temperatures*

Figures 2 and 3 give the results for forecasting minimum and maximum temperatures. For minimum temperatures, persistence shows only small variations from station to station except at Toledo, which had higher errors than its neighbors. For maximum temperature persistence there is a north–south trend of about 1.5°C across the forecast region, indicating more day-to-day variability in the southern part of the region. Again, Toledo had the highest error. Stations near Lake Michigan had reduced persistence variability, presumably due to the ameliorating effects of the lake.

Guidance forecasts for minimum temperatures showed more skill in the southern part of the region than in the northern part. The greatest improvements were at Grand Rapids and Flint, and the worst at Alpena, Traverse City and Sault Ste. Marie. For maximum temperatures, the highest skill scores were along a NNE–SSW line across the region. The greatest im-

TABLE 1. Forecast results for the Michigan area—Fall 1980. For temperature, the total error is in terms of a root-mean-square error. For precipitation, the total error is a mean Probability Score. See the text for more details.

			Contribution	
	Forecast interval (GMT)	Total error	Synoptic	Mesoscale
0000–1200	Minimum temperatures (°C)			
	climatology	4.5	3.9	2.2
	persistence	5.3	4.7	2.5
	NWS guidance	2.8	2.1	1.8
	consensus	2.4	1.7	1.7
1200–2400	Maximum temperatures (°C)			
	climatology	4.8	4.4	2.0
	persistence	5.4	4.7	2.4
	NWS guidance	2.8	2.1	1.7
	consensus	2.5	1.8	1.7
0000–1200	Precipitation probability (Probability Score)			
	climatology	0.190	0.111	0.079
	persistence	0.289	0.155	0.139
	NWS guidance	0.130	0.051	0.079
	consensus	0.126	0.051	0.075
1200–2400	Precipitation probability (Probability Score)			
	climatology	0.198	0.110	0.088
	persistence	0.369	0.194	0.175
	NWS guidance	0.130	0.045	0.085
	consensus	0.126	0.038	0.088

provement was at South Bend, and the worst at Sault Ste. Marie, Traverse City and Muskegon. Interestingly, the guidance forecasts in general have higher skill scores at stations in higher-population areas. Possibly, the weather is more consistent and predictable in these higher-population areas. More likely, the weather observations are more reliable at the larger stations, the result of better trained personnel and better equipment.

Consensus' improvement over guidance was positive at all but Muskegon for minimum temperatures. The greatest improvements were at Lansing and Detroit for minimum and maximum temperatures, respectively, and the worst were at Muskegon and Alpena.

As shown in Table 1 for both maximum and minimum temperatures, NWS guidance and consensus did considerably better in total rms error than persistence and climatology. For instance, NWS guidance had lower errors than persistence by 2.5°C for minimum temperatures and 2.6°C for maximum temperatures. Most of the improvements of NWS guidance and consensus over persistence are in the synoptic-scale contribution. For the mesoscale contribution, the improvement of NWS guidance and consensus over climatology and persistence was much less than on the synoptic scale. On the mesoscale, NWS guidance improved upon persistence by 0.7°C for both minimum and maximum temperatures.

Consensus improved over NWS guidance by 0.4°C for minimum temperatures and 0.3°C for maximum temperatures. Almost all of this improvement can be traced to the synoptic scale with practically no improvement on the mesoscale for either minimum or maximum temperatures.

b. Precipitation probability

Figures 4 and 5 give the results for probability of precipitation. In the first period for persistence there were low errors in the southwest and northeast sections with a band of higher values running from Traverse City to Toledo. The second period had a north–south trend, with a minimum at South Bend and a maximum at Sault Ste. Marie.

There were marked differences in the guidance skill scores in the first period, ranging from 0.38 at South Bend and 0.39 Sault Ste. Marie to 0.67 at Traverse City. In the second period, there was substantially less variation across the network.

In the first period, consensus showed some improvement over guidance in the southern part of the network, but negative values at Traverse City and Alpena. In the second period consensus improved somewhat over guidance at seven stations, but did worse than guidance at Alpena, South Bend and Toledo.

As shown in Table 1, NWS guidance did considerably better in total Probability Score than persistence and climatology. Significant gains over climatology and persistence are found on the synoptic scale. However, on the mesoscale there was almost no improvement of guidance over climatology (although a substantial improvement over persistence).

In total error, consensus was slightly better than NWS guidance in both periods. However, the improvement is probably too small to be significant and indicates there was no real ability to improve upon guidance on either the mesoscale or the synoptic scale. NWS guidance and consensus both had mesoscale contributions that were clearly greater than the synoptic scale contributions. Thus, unlike maximum/minimum temperatures, where the mesoscale and synoptic scale contributions are roughly equal, the current inaccuracy in precipitation probability results more from mesoscale variations and less from synoptic scale.

5. Conclusions

When the author started this experiment he expected that the forecasters would show little improvement over guidance on the synoptic scale, but at least some on the mesoscale. The results of this experiment were the opposite. Why is there no subjective forecasting ability on the mesoscale? It is difficult to interpret the collective thinking of all the forecasters that make up the consensus forecast. However, several factors may be important. Experiences of the author through the years have indicated that often mesoscale events which happen cannot be clearly defined in terms of "classical" subjective models. For instance, one morning Lansing would have a minimum temperature much below the surrounding stations in lower Michigan. The next time, however, under similar conditions, Detroit would be the cold spot. Some mesoscale phenomena apparently are extremely sensitive to the synoptic flow patterns that produce them. Small changes in the synoptic pattern can cause large changes in the mesoscale patterns. Synoptic-scale forecasts have to be accurate before the mesoscale forecasts will be reliable.

Consensus' improvement over guidance was 17% and 11% for minimum and maximum temperature forecasts, respectively, and 3% for the probability of precipitation. For temperature, Zurndorfer et al. (1979) reported that the local forecasters of the National Weather Service had an overall improvement over guidance ranging from 6% to 18% in the period 1973 through 1976. For precipitation, Charba and Klein (1980) reported an improvement of around 10% in the period 1976–78 (and decreasing with time). Thus, even though the overall experience levels of the forecasters in this experiment were low, there is apparently a surprisingly high level of improvement over guidance, especially for temperature. How can this be explained?

Gedzelman (1978) proposes that forecasting skill is mostly acquired in the first 30 forecasts. Since most of the participants in this study had made less than 30 forecasts, forecasting skill apparently is indeed acquired quickly. With MOS guidance available (and the use of it encouraged), the forecasters were modifying a forecast that was already usually very good. Since almost all of the improvement over guidance can be traced to the synoptic scale, the improvements were from modifications made to MOS for the stations as a whole.

Firestone (1979) suggested that synoptic-scale meteorological wisdom is gained faster than local experience. Gedzelman (1979) believes that forecasting skill is acquired rapidly at both the synoptic and mesoscales. The spatial analyses in Figs. 2–5 indicate that the stations with the best improvements were either in the area surrounding Ann Arbor where the forecasts were made (just west of Detroit), or at least were generally highest in southern lower Michigan. This tendency for the forecasts to be best locally possibly can be explained by the following factors. First, the forecasters probably had a better "feeling" and interest in the local area and had more specific weather information, possibly even by simply looking out a window. Second, perhaps ten stations were too many to cover in the time allotted for making the forecasts, and the most time and effort were expended on the nearest stations. This spatial variation would appear as a mesoscale contribution. Since overall consensus showed no mesoscale improvement over guidance; apparently, what was gained for nearby stations was lost at the stations farther away.

The results of this study support the findings of Baker (1982) that there are still substantial synoptic-scale errors remaining for temperature forecasting (approximately equal to the mesoscale). For precipitation, clearly there is more remaining error on the mesoscale, but the synoptic scale is still important. Hence, for precipitation the proposal that the remaining errors come more from the mesoscale than the synoptic scale appears to be valid, at least for the fall season. Seasonal variations undoubtedly exist.

On the mesoscale, guidance showed no improvement over climatology. This lack of improvement suggests that guidance contributed no further mesoscale information beyond climatology. Although substantial gradients in NWS guidance across lower Michigan are seen in precipitation probability, it appears that either the accuracy on the mesoscale is not any better than the normal climatological gradients, or that such gradients occur too infrequently to be important in the verification.

The results of this study suggest that zone forecasts probably contain a modest amount of mesoscale information. The atmospheric behavior on the synoptic scale is already very complex. The complexity undoubtedly increases at smaller and smaller scales. It is unreasonable to expect much improved mesoscale subjective forecasting without substantial computational help for the forecasters. For mesoscale phenomena that move, regional physical numerical models would be needed for projections beyond a few hours. For stationary phenomena related to specific land features, boundary layer models are needed. Specifically in the Great Lakes area, the mesoscale phenomena which might be amenable to modeling include 1) Great Lakes effects, 2) urban effects and 3) local very low-level orographic effects. In addition to providing new sources of mesoscale information, the National Weather Service is planning to upgrade weather communications and computational facilities at the local forecast offices. Hopefully, these local computational facilities will be able to provide the capability for making mesoscale model forecasts.

Acknowledgments. In addition to the author, the major participants in this study were J. Behr, P. Bodner, T. Hoogland, K. Jorae, M. Maguire, J. Petraska, P. Stamus and G. Weber.

REFERENCES

Baker, D. G., 1982: Synoptic-scale and mesoscale contributions to objective operational maximum/minimum temperature forecast errors. *Mon. Wea. Rev.*, **110**, 163–169.

Brier, G. W., 1950: Verification of forecast expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Bosart, L. F., 1975: SUNYA experimental results in forecasting daily temperature and precipitation. *Mon. Wea. Rev.*, **103**, 1013–1020.

——. 1983: An update on trends in skill of daily forecasts of temperature and precipitation at the State University of New York at Albany. *Bull. Amer. Meteor. Soc.*, **64**, 346–354.

Charba, J. P., and W. H. Klein, 1980: Skill in precipitation forecasting in the National Weather Service. *Bull. Amer. Meteor. Soc.*, **61**, 1546–1555.

Firestone, J. K., 1979: Comment on "Forecasting skill of beginners". *Bull. Amer. Meteor. Soc.*, **60**, 1206–1207.

Fritsch, J. M., and C. W. Kreitzberg, 1978: Workshop on local weather forecast techniques—1980 and beyond, Drexel University, Philadelphia, Pa., 10–11 March 1977. *Bull. Amer. Meteor. Soc.*, **59**, 293–296.

Fujita, T. T., 1981: Tornadoes and downbursts in the context of generalized planetary scales. *J. Atmos. Sci.*, **38**, 1511–1534.

Gedzelman, S. D., 1978: Forecasting skill of beginners. *Bull. Amer. Meteor. Soc.*, **59**, 1305–1309.

——. 1979: Response to comments on "Forecasting skill of beginners". *Bull. Amer. Meteor. Soc.*, **60**, 1208–1209.

——. 1981: Limits on forecast accuracy due to small-scale weather variations. *Bull. Amer. Meteor. Soc.*, **62**, 1570–1576.

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.

Jorgensen, D. L., 1967: Climatological probabilities of precipitation for the conterminious United States. NOAA Tech. Rep. WB-5.

National Climatic Center, 1973: *Climatology of the United States*, No. 84 [National Climatic Center, Federal Building, Asheville, N.C.].

National Weather Service, 1985: Operations of the National Weather Service. 251 pp. [available from the National Weather Service, Silver Spring, MD].

Ramage, C. S., 1976: Prognosis for weather forecasting. *Bull. Amer. Meteor. Soc.*, **57**, 4–10.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171–1179.

Zurndorfer, E. A., J. R. Bocchieri, G. M. Carter, J. P. Dallavalle, D. B. Gilhousen, K. F. Hebenstreit and D. J. Vercelli, 1979: Trends in comparative verification scores for guidance and local aviation/public weather forecasts. *Mon. Wea. Rev.*, **107**, 799–811.