Dr. Glahn

U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
SYSTEMS DEVELOPMENT OFFICE
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 83-10

A COMPARATIVE VERIFICATION OF GEM AND MOS

Thomas J. Perrone and Robert G. Miller

July 1983

A COMPARATIVE VERIFICATION OF GEM AND MOS

Thomas J. Perrone and Robert G. Miller

1. INTRODUCTION

GEM is an acronym for a statistical weather forecasting technique which predicts the probability distribution of all surface weather elements hour by hour. GEM uses only the current local surface weather conditions as predictors. Climatological information is also used, in two ways: implicitly, through the Regression Estimation of Event Probabilities (REEP) (see Miller, 1964) in the GEM process, and explicitly, to supply location-specific information to the GEM forecast. From the distribution of probabilities of the forecasted weather events, GEM also makes categorical predictions.

"G" indicates generalized. The same statistical equations can be applied at any location and for any time period. "E" stands for equivalent, because of GEM's equivalence (as a linear approximation) to a Markov chain. "E" also stands for exponential, a characteristic of the particular form of the Markov process necessary to model events which occur in continuous time. "M" indicates that the technique is a Markov process.

An excellent definition of a Markov process as applied to a physical situation, such as weather forecasting, is given by Feller (1950):

> "In stochastic processes the future is never uniquely determined, but we have at least probability relations enabling us to make predictions.... The term 'Markov process' is applied to a very large and important class of stochastic processes.... Conceptually, a Markov process is the probabilistic analogue of the processes of classical mechanics, where the future development is completely determined by the present state and is independent of the way in which the present state has developed...in contrast to processes...where the whole past history of the system influences its future."

The motivation for GEM's development is the need to provide accurate, yet computationally feasible, computer-generated short-term weather forecasting guidance based on the very latest weather information. In general, persistence, though essentially a "no-skill" technique, has been the most skillful guidance available for forecasts of most weather elements for projections ranging from 0- to 6-hours.

Model Output Statistics (MOS) (see Glahn and Lowry, 1972) is now widely accepted as a highly-skilled purveyor of statistical-dynamic weather forecasting guidance. The input to MOS requires data from models, which, however, results in a gap of about 5 hours between upper air observations (about 2 hours from surface observations) and the availability of MOS. The gap results from a combination of two factors: first, the amount of centralized computer time necessary to generate the model output and, in turn, the MOS forecasts; and second, dynamic model instability within the first twelve hours which renders somewhat dubious much of the model output valid for the first 6 hours of the model run.

In the AFOS (Automation of Field Operations and Services) era, increased emphasis is being placed upon locally-generated computer weather forecasting techniques. Forecasting systems which operate skillfully within the computer time and space constraints imposed by AFOS will likely be favored for implementation.

GEM incorporates all the information contained within the surface observation. GEM also models very well the effect of persistence. Because GEM uses only the observatonal elements as predictors (modified by local climatology) together with a highly sophisticated computation scheme, it is well-suited to on-site computer implementation.

The philosophy underlying GEM is rooted in the writings and lectures of Wiener (1948, 1950, 1956). There, he makes the case for a probabilistic approach to meteorological prediction and for a linear solution to the problem. The first articulation of GEM was in 1964, in a proposal to the U.S. Air Force's Air Weather Service (AWS) to incorporate specials and other irregularly observed weather conditions such as those in pilot reports, radar observations, and satellite images (Miller, 1968). Although not undertaken then, the first implementation of GEM did occur in 1977 at St. Louis University, in conjunction with AWS (Miller et al., 1977). Subsequent refinements relating to GEM have been reported by Miller (1979) and Whiton (1977).

To date, the most comprehensive study involving GEM can be found in Miller (1981). That study describes the development of GEM within the National Weather Service and documents some of the experiments run to verify GEM against persistence. Those interested in acquiring an in-depth understanding of GEM, particulary its statistical aspects, should consult that work. To gain some understanding of the comprehensiveness of GEM, consult Fig. 1, a map of stations whose data were used in the development of GEM.

This study accomplishes two purposes:

a. After presenting some background material necessary for understanding the results of this study, and defining its scope (Section 2), Section 3 reports on an extensive statistical verification experiment of GEM vs. MOS.

b. It describes some statistical experiments carried out subsequent to those reported previously, and motivated by the GEM-MOS comparative verification.

This study, however, does not document the extensive and successful verification of GEM against persistence. Details of this verification may be found in Miller (1981) and Miller et al. (1981).

Section 4 reports on the results of experiments to blend GEM and MOS for ceiling and visibility forecasting. The motivation for this effort was to create a forecast system that capitalizes on GEM's short-range forecast capabilities together with MOS's longer-range abilities. Section 4 also presents a conceptual cost-benefit analysis of blending.

Section 5 describes a series of experiments using feedback, in an effort to improve both GEM and MOS temperature forecasting capabilities. Section 6 discusses three miscellaneous topics that were subjects of ancillary statistical experiments. Among the topics discussed are:

a. Use of a collection of multivariate statistical techniques, somewhat related to one another, to make categorical forecasts.

b. Use of blends of local monthly climatology and local hourly climatology to attempt to account more fully for local station effects than by use of local hourly climatology alone.

c. Use of a variation of GEM's P-star process, termed "unaccumulated P-stars", to make categorical forecasts.

Section 7 ends the body of this report with conclusions and some relevant remarks. Terms such as "P-star" have special meaning with respect to GEM. Definitions of such terms, unique to GEM, or which have special usage, are given in Appendix A. In meteorological verification, a number of standard scores are routinely used to report verification results. Among those used in this study are the Heidke Skill Score, percent correctly forecast, threat score (for ceiling and visibility) a chi-square goodness-of-fit measure on margins, mean absolute error (for temperature), mean algebaric error (also for temperature), number of "large" errors (also for temperature), and the Brier Score (for ceiling, visibility, and total cloud amount). These scores are also defined in Appendix A.

2. SCOPE OF STUDY

This section presents background information necessary for understanding the MOS-GEM comparative verification presented in Section 3. It also discusses the scope of the study.

The MOS forecast system has undergone extensive development. Verification of MOS forecasts against both observed conditions and on-station forecaster performance has also been thoroughly documented (see, e.g. Carter, Bocchieri, Dallavalle (1982)). The purpose of the next section is to present comprehensive verification results of GEM against the known guidance standard, MOS.

The general guidelines for the study were to verify weather elements common to both GEM and MOS, for categories of these elements compatible with the two systems, and for projections when the products of both forecast systems were available. The weather elements in common are ceiling, visibility, total cloud amount, temperature, dew-point depression, and wind.

Here are the ways each element was verified:

a. Ceiling. Ceiling was verified as a categorical weather element in six categories. The definition of each category is given in Table 1.

Scores for the ceiling comprise percent correctly forecast, Heidke Skill Score, a chi-square goodness-of-fit measure on marginals, threat score on the lowest two categories combined, and the Brier Score calculated from the probabilities associated with each ceiling category (see Appendix A for definitions of these scores).

b. Visibility. Visibility was also verified as a categorical weather element in six categories. The categories are given in Table 2.

Scores for visibility include percent correctly forecast, Heidke Skill Score, a chi-square goodness-of-fit measure on marginals, threat score on the lowest two categories, and the Brier Score calculated from the probabilities associated with each visibility category.

c. Total Cloud Amount. Total cloud amount was verified as a categorical weather element. The categories correspond to the well-known sky conditions CLEAR, SCATTERED, BROKEN, and OVERCAST. There is, however, a small difference in the definition of these terms as used in MOS and GEM. MOS forecasts the total cloud amount as total opaque (non-thin) sky cover, while GEM forecasts the total could amount as the total sky cover without regard to opaqueness. For the most part, the total sky condition will be classified the same under either defintion. Occasionally, however, when significant amounts of thin (non-opaque) clouds are present, the total sky condition will differ under the two definitions. The MOS-GEM comparative verification does not attempt to correct for the difference in the definitions, so the verification results should be viewed with some caution. Also, as a consequence of the definitional differences, the sample sizes used for GEM and MOS differ slightly. Scores for the total cloud amount are percent correctly forecast, Heidke Skill Score, a chi-square goodness-of-fit measure on marginals, and the Brier Score calculated from the probabilities associated with each total cloud amount category.

d. Temperature. Temperature was verified as both a continuous and a categorical weather element. When verified as a continuous weather element, scores included the mean absolute error, the mean algebraic error, and a count of the number of "big" errors (number of errors between forecast and observed temperatures of $\pm$ 10°F or more). (See Appendix A for more details on these scores.) When verified as a categorical weather element, temperature was divided into 30 categories. The temperature category definitions are given in Table 3.

Scores for temperature as a categorical weather element are percent correctly forecast, Heidke Skill Score, and a chi-square goodness-of-fit measure on the marginals.

e. Dewpoint Depression. Dewpoint depression was verified as a categorical weather element, whose categories are given in Table 4.

Scores for dewpoint depression include percent correctly forecast, Heidke Skill Score, and a chi-square goodness-of-fit on marginals.

f. Wind. Wind was verified as a categorical element. The categories, combinations of wind speeds and directions, are given in Table 5.

Scores for wind include percent correctly forecast, Heidke Skill Score, and a chi-square goodness-of-fit on marginals.

The MOS forecast data were taken from archives, as were the verifying observations, while the GEM forecasts were made afresh for the study. Since

GEM needs only an observation as input, and can make a forecast for any projection, availability of GEM forecasts were dependent only on the availability of initial observations. Within the general scope of the study, MOS archived forecasts were available for projections of 6-, 9-, 12-, 15-, 18-, 21-, 24-, 27-, and 30-hours for temperature and dewpoint depression, and at 6-, 12-, 18-, 24-, and 30-hours for the other elements. Accordingly, verification projections picked for the study were those for which MOS data were available.

Some classification terminology about forecast projections are necessary because MOS and GEM define projections differently. In MOS, the projections are reckoned from the time of the model run which produces the model output predictors (00 GMT or 12 GMT); the observations used by MOS as predictors are usually available 3 hours later (03 GMT or 15 GMT). Sometimes 02 GMT or 14 GMT observations are used, but when no observations are available, MOS uses "backup" equations, which use only model output as predictors. No attempt was made in this study to differentiate among MOS forecasts made with 03 GMT (15 GMT), 02 GMT (14 GMT), or no observational predictors. We took MOS as we found it in the archives, much as it would be available in a real-time setting.

GEM projections are reckoned from the time of the observation used as a predictor. In this study, forecast projections are defined as they are used with GEM. GEM and MOS were comparatively verified in three modes: scientific, operational, and special operational.

In the scientific mode, MOS and GEM share the same observation as input. In the operational mode, the GEM observation used is 6 hours later than the observation used by MOS. In the special observational mode, the GEM observation used is 12 hours later than that used by MOS. The scientific and operational modes are illustrated in the "time lines" of Fig. 2. This figure shows, for these modes, the relative times of the dynamic model run (labeled LFM, for Limited Fine Mesh model), MOS and GEM observation times, and the verification times for 3- and 9-h projections.

The motivation for employing three projection modes was to fully test the validity of GEM from differing viewpoints:

a. The scientific comparison, as its name implies, is a "pure" comparison of GEM and MOS as statistical forecasting techniques. It measures the extent to which GEM, a "classical" (i.e. non-dynamic) statistical technique, which is only limitedly station-specific, can compete against MOS, which not only has model predictors as input but is developed in 6-month seasons for individual stations or small sections of the country.

b. The operational comparison tests GEM's capabilities to operate with later data than centrally-produced MOS and helps to evaluate GEM's usefulness for aviation FT preparation. In this study, GEM forecasts using observations at 09 GMT and 21 GMT simulates production of forecast guidance for 0940 GMT and 2140 GMT FT file times.

c. The special operational comparison tests GEM's capabilities vs MOS's during the periods 00-04 GMT and 12-16 GMT, when the only MOS guidance available is that derived from the previous model run cycle. In this study, GEM forecasts from observations at 15 GMT simulate guidance for 1540 GMT FT file times.

The data used in this study were for a full year (a warm season and a cool one)* from April 1, 1980 through March 31, 1981. The 21 stations used in the comparative verification are indicated in Fig. 3 and listed in Table 6. Stations DCA, PHL, MIA, MSY, MSP, SLC, and SFO (indicated on Fig. 3 by five-pointed stars) were also used in an earlier comparative verification of GEM vs persistence (see Miller, 1981 and Miller et al., 1981).

As reported in the next section, the GEM-MOS comparative verification was carried out for 3- and 6-h projections for temperature and dewpoint depression, and 3- and 9-h projections for the remaining elements. It was clear from the inception of the study that at large projections, MOS would likely be superior to GEM because of the presence of model output as predictors. Part of the motivation for the current study is to determine a crossover point in time where GEM's superiority, if any, is supplanted by that of MOS.

Although our data base possesses the ability to support a comparative verification out to 30 hours, prudent use of personnel and computer resourses indicated limiting our study to a smaller number of projections. Sufficient information is available from examining projections no greater than 9 hours to determine in a general way the MOS-GEM crossover points for each element. Any finer resolution of the crossover than that determined in this study, is prevented by the inability to obtain archived MOS forecasts spaced closer together than 3 hours for temperature and dewpoint depression, and 6 hours for the other elements.

3. GEM-MOS COMPARATIVE VERIFICATION

This section presents detailed results of a GEM-MOS comparative verification, for the elements of ceiling, visibility, total cloud amount, temperature, dewpoint depression, and wind. The main method used to display results is presentation of the scores applicable to the particular element in tabular form, stratified by season and time of the GEM input observation used, as well as in aggregated form. Since the sample sizes of the data in aggregated form are larger than for any of the stratified tables, greater confidence may be placed in the resulting aggregated table results. Consequently, aggregated table results are used to support our determination of the GEM-MOS crossover. We also note, as appropriate, any patterns that arise in the stratified results. Where appropriate, we display contingency tables to aid in conveying the flavor of results and to illustrate certain strengths or weaknesses of the forecast processes. Due, however, to the overwhelmingly large amount of data processed for this study, display of data in contingency table form is kept to a minimum. Detailed contingency table data are available from the authors. Chi-square measures are available only for the stratifications.

A. Ceiling

For the element ceiling at a 3-h projection under the operational comparison, GEM in the aggregate is superior to MOS in percent correct, Brier

*For definition of "warm season" and "cool season", See Appendix A.

Score, Heidke Skill Score, and threat score. Under the scientific comparison, GEM in the aggregate is slightly superior to MOS in percent correct, but MOS is favored for the remainder of the scores (See Table 7).

Among the stratified results, the same conclusions hold as for the aggregated results, except in the scientific comparison: GEM and MOS tie in Brier Score for the warm and cool season/03 GMT GEM input time stratification, GEM out-performs MOS in percent correct for the 03 GMT GEM input time stratifications (regardless of season), and GEM has a better threat score for the warm season while MOS is better for the cool season for 03 GMT GEM input time stratification. MOS is better on all scores for 15 GMT GEM input time stratification. MOS chi-squares are preferred over GEM throughout Table 7.

For ceiling at a 9-h projection, for both the operational or scientific comparsion, MOS is favored over GEM in the aggregate for all scores except under the operational comparison for the threat score, where GEM is superior (see Table 8). Among the stratified results, MOS outperforms GEM on all stratifications, except that GEM is superior on the threat score for all operational comparisons and for the cool season/15 GMT GEM input time stratification under the scientific comparison. GEM also outperforms MOS in percent correct for the warm season/21 GMT GEM input stratification under the operational comparison, and for the warm season/03 GMT GEM input stratification under the scientific comparison. MOS chi-squares are smaller than for GEM throughout Table 8, except for the warm season/15 GMT GEM input time stratification under the scientific comparison. For the chi-square measure, smaller is better.

Among the special operational comparisons for ceiling, in the aggregate, GEM is superior to MOS for the 3-h projection; MOS is favored for the 9-h projection except for the threat score, where GEM is superior (see Table 9). Among the stratified results, the same conclusions hold, except for the 9-h projection: GEM is superior in percent correct for the warm season/03 GMT GEM input time stratification, and the two forecast processes tie in percent correct for the warm season/15 GMT GEM input time. Also in the 9-h projection stratification results, GEM's threat score is better in the cool season (regardless of GEM input time), while MOS's threat score is better in the warm season. MOS chi-squares are smaller than for GEM throughout Table 9, except for the cool season/03 GMT GEM input time and warm season/15 GMT GEM input time stratifications for the 9-h projection.

For the operationally critical ceiling categories 1 and 2 (ceilings less than 500 ft), GEM at 3 hours, under the operational comparison, produces 108 "hits" (number correct) for 240 forecasts in the two lowest ceiling categories, while MOS achieves 55 "hits" for 342 forecasts. GEM, therefore, achieves 53 more hits with 102 fewer forecasts than MOS. At 9 hours, also under the operational comparison, GEM produces 38 "hits" with 251 forecasts, while MOS produces 37 "hits" with 219 forecasts. At 9 hours, GEM achieves only one more hit at the cost of 42 additional forecasts than MOS (see Table 10).

In addition to the scores already discussed, Table 10 also displays the biases for each ceiling category. (For definition of bias, see Appendix A.) For the lowest two categories of ceiling, the GEM biases are below one, while MOS is above one, for the 3-h projection. For the 9-h projection, the biases of the first two categories are slightly below one for both of the lowest two MOS categories and for GEM ceiling category 2; the bias, however, is substantially above one for GEM ceiling category 1.

B. Visibility

For the element visibility at a 3-h projection, under the operational comparison, GEM in the aggregate is superior to MOS on all scores. Under the scientific comparison GEM, in the aggregate, is superior to MOS in percent correct and Heidke Skill Score, but MOS is favored for the Brier and threat scores (see Table 11). Among the stratified results, GEM is superior to MOS on all scores under the operational comparison, except that MOS and GEM have the same Brier Score for the warm season/09 GMT GEM input time stratification.

Under the scientific comparison, the stratification results reflect the somewhat mixed aggregated results. In the Brier Score, GEM and MOS tie for the warm season/03 GMT GEM input time stratification, GEM is slightly superior for the cool season/03 GMT GEM input time stratification, while MOS is favored for both seasons for GEM input time of 15 GMT. For the percent correct measure, GEM is superior to MOS for both seasons stratified by 03 GMT GEM input time, while MOS is favored for both seasons stratified by 15 GMT GEM input time. Among the stratified Heidke Skill Score results, GEM is superior except for the warm season/15 GMT GEM input time stratification. The stratified threat score results are also mixed: MOS is favored for both the warm season/03 GMT GEM input time and cool season/15 GMT time stratifications. Both forecast procedures possess a no-skill threat score for the warm season/15 GMT GEM input time stratification, while GEM is favored for the cool season/03 GMT GEM input time stratification. MOS chi-squares are smaller than for GEM throughout Table 11, except for cool season/21 GMT GEM input time stratification, in which the GEM chi-squares are smaller.

For the visibility at a 9-h projection, MOS, in the aggregate, is superior to GEM for all scores under both the operational and scientific comparisons (see Table 12). Under both comparisons within each of the stratifications, each Brier and Heidke Skill Score favors MOS. Under the operational comparison, the percent correct is greater for MOS in the 09 GMT GEM input stratifications (regardless of season), and greater for GEM in the 21 GMT GEM input stratifications.

Under the operational comparison among the stratifications the threat score is mixed. GEM is favored for the cool season/21 GMT GEM input stratification, MOS is favored for the warm season/21 GMT GEM input and cool season/09 GMT GEM input stratifications, while MOS and GEM both possess no-skill in the threat score for the warm season/09 GMT GEM input time stratification.

Under the scientific comparison MOS attains a greater percentage correct for all stratifications except the warm season/15 GMT GEM input stratifictaion, where there is a tie. Under the scientific comparison, among the stratifications, MOS also attains a higher threat score than GEM except for a no-skill tie for the warm season/15 GMT GEM input stratification. MOS chi-squares are smaller than GEM throughout Table 12.

Under the special operational comparison, for the 3-h projection, GEM is superior, in the aggregate, on all measures (see Table 13). For the 9-h projection, MOS, in the aggregate, is favored on all measures except the threat score, in which GEM is superior.

For the 3-h projection the aggregated outcomes are also true for the stratifications, except for a no-skill threat score tie between GEM and MOS for the warm season/15 GMT GEM input stratification. For the 9-h projection, the aggregated outcomes hold for the stratifications with these exceptions: GEM is favored with a higher percent correct than MOS in the warm season/15 GMT GEM input stratification, and there is a no-skill tie between GEM and MOS in the warm season/15 GMT GEM input stratification for threat score--otherwise MOS is superior. MOS chi-squares are smaller throughout Table 13.

For the operationally critical visibility categories 1 and 2 (visibilities less than 3 miles), GEM at 3 hours, under the operational comparison, produces 63 "hits" (number correct) for 151 forecasts of categories 1 and 2, while MOS achieves 35 "hits" for 268 forecasts. GEM, therefore, achieves 28 more hits with 117 fewer forecasts than MOS. At 9 hours, also under the operational comparison, GEM produces 23 "hits" with 158 forecasts, while MOS produces 30 "hits" with 160 forecasts. At 9 hours, GEM produces 7 fewer "hits" with 2 fewer forecasts than MOS (see Table 14). Table 14 also shows the biases for each visibility category. For the lowest two categories of visibility, the biases for GEM and MOS are greater than one, except for GEM at 3 hours, where the biases of the lowest two categories are below one.

C. Total Cloud Amount

For the element total cloud amount at a 3-h projection under the operational comparison, GEM in the aggregate is superior to MOS on all scores (see Table 15). Under the scientific comparison, the reverse is true.

Among the stratified results, GEM is superior for all stratifications of the operational comparison, except for a tied Brier Score with MOS for the warm season/09 GMT GEM input time stratification, and for some of the chi-square scores.

Under the scientific comparison, MOS is favored in the Brier Score for all stratifications. GEM achieves a higher percent correct for the cool season, MOS in the warm season. MOS is favored in the Heidke Skill Score for all stratifications, except for a GEM-MOS tie in the cool season/15 GMT GEM input time stratification.

Throughout Table 15, GEM chi-squares are larger than those of MOS, except for these stratifications: under the opertional comparison, cool season/21 GMT GEM input time; under the scientific comparison, warm season/03 GMT GEM input and cool season/15 GMT GEM input stratifications.

For a 9-h projection under both the operational and scientific comparisons, MOS is favored over GEM for those scores reported in the aggregate (see Table 16).

Under the operational comparison, among the stratifications, MOS is favored over GEM on all measures, except for certain chi-square measures; a tie in the

Brier Score in the warm season/09 GMT GEM input time stratification; and in the percent correct for the cool season/21 GMT GEM input time stratification, in which GEM is favored.

Under scientific comparison, MOS is favored over GEM for all stratifications in the percent correct, Brier Score, and Heidke Skill Score.

Throughout Table 16 GEM chi-squares are in general larger than those of MOS, except for the cool season/09 GMT GEM input time stratification under the operational stratification and the cool season/15 GMT GEM input time stratifiction under the scientific comparison.

Turning to the special operational comparison, for the 3-h projection, in the aggregate and for each of the stratifications, GEM is superior to MOS in percent correct, Brier Score and Heidke Skill Score (see Table 17). For the 9-h projection, however, in the aggregate MOS is superior to GEM for the same three scores.

Of the 9-h projection stratifications, MOS is favored over GEM for each Brier Score. Also, MOS is favored over GEM for the percent correct and the Heidke Skill Score, except for the cool season/03 GMT GEM input time stratification.

Throughout Table 17, GEM's chi-squares are larger than those of MOS except for the cool season/15 GMT GEM input time stratifications for both the 3- and 9-h projections.

D. Temperature

For the element temperature at a 3-h projection using the operational comparison, GEM in the aggregate is superior to MOS on all scores reported in the aggregate, except for the mean algebraic error, which favors MOS (see Table 18). For the scientific comparison in the aggregate, MOS is favored over GEM for each of the scores reported in the aggregate.

Among the stratifications, under the operational comparison, GEM achieves a lower mean alsolute error for each stratification except the cool season/21 GMT GEM input time stratification. The magnitude of the mean algebraic error among the stratifications is smaller for GEM in the cool season/09 GMT GEM input stratification and smaller for MOS in the warm season/21 GMT GEM input stratification. For the other two stratifications, comparison of the mean algebraic error are somewhat indeterminate: the scores achieved by each forecast process is of about equal magnitude but of opposite sign. Among the stratifications, each number-of-large-error score favors GEM. Among the stratifications' percent correct and Heidke Skill Score measures, GEM's scores are higher for the 09 GMT GEM input time stratification (regardless of season), while MOS's scores are higher for the 21 GMT GEM input time stratifications.

Of the scientific comparison stratifications, in each stratification MOS is favored for the mean absolute error, number of large errors, percent correct, and Heidke Skill Score. For the mean algebraic error, the magnitude of GEM's error is lower in the warm season, while MOS's error is lower for the cool season/15 and 21 GMT GEM input time stratification. For the cool season/03 GMT GEM input time stratification, the results are indeterminate, as the scores for each forecast process are of similiar magnitude but of opposite sign.

Throughout Table 18 GEM chi-squares generally are larger than those for MOS, except for the cool season/09 GMT GEM input time stratification under the operational comparison.

For temperature at a 6-h projection, using either the operational or scientific comparisons, MOS is generally favored over GEM for each measure, whether viewed in the aggregate or for each of the stratifications (see Table 19). The only exceptions under both the operational and scientific comparisons, are for the mean algebraic error, in which GEM is superior for the warm season/15 GMT and 21 GMT GEM input time stratifications. Throughout Table 19, MOS chi-squares are smaller than those of GEM.

Turning to the special operational comparison, for the 3-h projection, GEM, in the aggregate, is superior for the mean absolute error, number of large errors, percent correct, and Heidke Skill Score. MOS, however, achieves in the aggregate a smaller mean algebraic error (see Table 20).

Among the stratifications for the 3-h projection, for mean absolute error, number of large errors, and percent correct, GEM is superior to MOS for each stratification except that of the cool season/15 GMT GEM input time. GEM's Heidke Skill Score is superior for the 03 GMT GEM input time stratifications (regardless of season) while MOS is favored for the warm season/15 GMT GEM input time stratification. The magnitude of the MOS mean algebraic error for the cool season stratifications (regardless of GEM input time) is lower than for GEM, while the warm season algebraic error comparisons for the two forecast processes are indeterminate, because the scores are of similar magnitude but opposite sign.

For projections of 3 hours, the GEM chi-square measures are larger than for MOS for each of the stratifications, except for the cool season/03 GMT GEM input time stratification, where the GEM chi-square is smaller by 0.1.

The 3-h stratification results contrast somewhat with the aggregate results: in the aggregate GEM is generally superior to MOS, but MOS's performance is superior to GEM's on all measures for the cool season/15 GMT GEM input time stratification.

For 6-h projections, MOS is superior to GEM on all scores, both in the aggregate and for each stratification. The sole exception is the number of large errors for the cool season/03 GMT GEM input time stratification; GEM achieves one fewer number of large errors.

E. Dewpoint Depression

For the element dewpoint depression at a 3-h projection for the operational comparison, GEM, in the aggregate, is superior to MOS for the percent correct and Heidke Skill Score measures (see Table 21). The same result holds for the stratifications, except that MOS is favored on the percent correct measure for the cool season/21 GMT GEM input time stratification. For the scientific comparison MOS is favored over GEM, both in the aggregate and within each of the stratifications, for the percent correct and Heidke Skill Score.

Throughout Table 21, MOS achieves lower chi-square values for each stratification, except for the cool season/09 GMT GEM input time and warm season/21 GMT GEM input time stratification under the operational comparison. Under the scientific comparison, the GEM chi-square for the cool season/15 GMT GEM input time stratification is 119 larger than the corresponding MOS chi-square value.

For dewpoint depression at a 6-h projection, for both the scientific and operational comparisons, in the aggregate and for each of the stratifications, MOS is superior to GEM in percent correct and the Heidke Skill Score (see Table 22). Throughout Table 22, GEM chi-squares are larger than those of MOS, except for the cool season/03 GMT GEM input stratification under the scientific comparison.

Some of the differences between the chi-square values achieved by the forecast processes are quite striking, and all favor MOS. Under the operational comparison, we note a difference of 329 for the cool season/09 GMT GEM input time stratification, 154 for the cool season/21 GMT GEM input time stratification, and 194 for the warm season/09 GMT input time stratification. Under the scientific comparison, the difference is 140 for cool season/15 GMT GEM input time stratification.

Turning to the special operational comparisons, for the 3-h projection GEM is superior to MOS, when considered in the aggregate and in each of the stratifications, for the percent correct and the Heidke Skill Score measures (see Table 23). GEM chi-square values are lower than for MOS in each stratification except the cool season/15 GMT GEM input time stratification.

For the 6-h projection MOS, in the aggregate, is favored over GEM in percent correct and for the Heidke Skill Score. These results hold as well for the stratifications, too, except for: the warm season/03 GMT GEM input time stratification for percent correct; and the warm season stratifications regardless of GEM input time, for the Heidke skill score.

For the 6-h projection MOS chi-squares are larger for 03 GMT GEM input stratifications (without regard to season), while GEM chi-squares are larger for the 15 GMT GEM input stratification. The GEM-MOS chi-square difference for the cool season/15 GMT GEM input time stratification is noteworthy: 102, in favor of MOS.

F. Wind

For the element wind at a 3-h projection, for the operational comparison, GEM in the aggregate is superior to MOS for percent correct and Heidke Skill Score measures (see Table 24). Among the stratifications, however, GEM is superior for the 09 GMT GEM input time stratifications (regardless of season), while MOS is favored for the 21 GMT GEM input time stratifications.

For the scientific comparison for the 3-h projection, MOS is superior, both in the aggregate and for each of the stratifications, for the percent correct and Heidke Skill Score measures.

Throughout Table 24 MOS chi-squares are smaller than those for GEM. The largest difference between the GEM and MOS chi-square values is a noteworthy 211, which favors MOS and occurs in the cool season/21 GMT stratification.

For wind at a 9-h projection, MOS is superior to GEM on all measures. MOS is favored in both the aggregate and among the stratifications for both the operational and scientific comparisons (See Table 25). Throughout Table 25 MOS chi-squares are very much smaller than for GEM. The largest difference is 767, favoring MOS, which occurs in the cool season/09 GMT GEM input time stratification under the operational comparison.

Turning to the special operational comparison for the 3-h projection, GEM in the aggregate is superior to MOS for the percent correct and Heidke Skill Score measures. This result also holds among the stratifications, except for the warm season/15 GMT GEM input time stratification, in which MOS is favored (see Table 26).

For the 9-h projection, MOS is superior to GEM for both percent correct and Heidke Skill Score in the aggregate as well as among the stratifications.

Throughout Table 26, MOS chi-squares are smaller than for GEM. The biggest noteworthy differences between MOS and GEM chi-square values (each favoring MOS) occur in these 9-h projection stratifications: cool season/15 GMT GEM input time (561), warm season/15 GMT GEM input time (412), and cool season/03 GMT GEM input time (209).

G. Summary

A summary of the salient results of the GEM-MOS comparative verification is displayed in Table 27. The table expresses the aggregated results in a fractional form. The number of scores favoring GEM forms the numerator; the total number of scores used for the particular weather element is the denominator. These "fractions" are displayed for the special operational, scientific, and operational comparisons, and for the two projections of each element. Since the chi-square measure is not available in aggregate form, it does not enter as one of the scores used in the "fractions". Displayed, however, in Table 27 are the number of stratifications for which the GEM chi-squares are less than those of MOS (a minimum of zero, maximum of four). Major differences in the aggregated results and the results among the stratifications are tagged and identified in footnotes to Table 27.

We may summarize the results in words this way: The order of elements listed in Table 27 is the order of greatest to least skill for GEM in comparison to MOS (i.e., ceiling is most skillful, wind, least skillful). A natural dividing point appears in the table between total cloud amount and temperature. The elements seem to fall into two groups comprising ceiling, visibility, and total cloud amount (elements of major interest for aviation forecasting) on the one hand and temperature, dewpoint depression, and wind on the other. The first group (major aviation elements) is marked by substantial

GEM superiority at 3 hours with no major differences between aggregate and stratified results. GEM's performance is strongest in the operational and special operational comparisons, and MOS is strongest in the scientific comparisons.

The scientific comparison performance of the total cloud amount is similar to that for temperature, dewpoint depression, and wind (lesser aviation elements). Total cloud amount behaves like the ceiling and visibility (major aviation elements), however, in that there are no major differences between aggregate and stratified results. Some of the difference that does exist in the performance of the total cloud amount in comparison with the other major aviation elements may be due to differences in the way GEM and MOS define total cloud amount (see Section 2). From the available information, grouping total cloud amount with ceiling and visibility seems justified.

Major differences occur between the aggregate and stratified results for temperature, dewpoint depression, and wind. These differences are indicated in the footnotes to Table 27. Briefly summarized, they are: for temperature, some evidence of MOS superiority in 3-h forecasts made during the MOS 12 GMT cycle; for dewpoint depression, some evidence of GEM superiority in 6-h forecasts made in the warm season from 03 GMT GEM input data; and for wind, some evidence of MOS superiority for certain of the 3-h stratified forecasts made under the special operational and operational comparisons.

GEM's superiority over MOS, where it occurs, has not been obtained without paying some sort of verification price. Throughout the whole verification scheme, GEM generally achieved higher (less desirable) chi-square scores, indicating "less good" fit with marginal expectations, than did MOS. Of the entries in Table 27, (where a maximum count of four is possible for the chi-square measure) the largest number of stratifications for which the GEM chi-square measure is less than that of MOS is three (dewpoint depression at 3 hours, special operational comparison). For dewpoint depression and wind, in five of the stratifications (out of a possible 24), the difference between GEM and MOS chi-square scores is quite large, exceeding 100. The reason for the generally larger GEM chi-squares may be that GEM is designed to forecast for all hours while MOS is hour-specific. Some of the lack of fit of GEM may be due to its use for only a subset of the hours for which it was derived.

Tables 28 and 29 express the results of the GEM-MOS comparative verification in a different form: determination of the "cross-over" in skill between the two techniques under operational conditions. Since GEM can be produced for any hour of the day from the latest available observation, while available MOS guidance uses observational predictors present only twice daily, Tables 28 and 29 indicate which forecast process is most accurate, conditioned on the time of the day and the forecast projection of interest. The circled items in both tables indicate where information from the study's results were used. The 03 GMT and 15 GMT data points represent results from the special operational comparison. The 09 GMT and 21 GMT data points represent results from the operational comparison. The scientific comparison results were used to "fill in" the crossover trends between 05 GMT and 09 GMT and between 17 GMT and 21 GMT.

It should be noted that the scientific comparison results cannot be used directly in these tables. MOS forecasts, though made with 03 GMT and 15 GMT observational predictors, are generally not available until sometime between 04-05 GMT and 16-17 GMT. The MOS-GEM results from the scientific comparisons are, therefore, adjusted to reflect:

a. Non-availability of new MOS guidance between 04-05 GMT and 16-17 GMT.

b. The deterioration of MOS forecasts relative to GEM's, when GEM uses later observations as input.

There is little difference in the results in both tables in the interval 09-16 GMT and 21-04 GMT, because the operational and special operational differences between MOS and GEM are similar.

4. BLENDING GEM AND MOS

This section presents results of a composite forecast system, derived by statistically blending GEM and MOS, and compares the results of the composite system against GEM and MOS singly. Mr. Joseph R. Bocchieri, formerly with Techniques Development Laboratory*, suggested the blending experiment (Bocchieri, 1982). In principle, our blending experiment is similar to one carried out by him for precipitation forecasting (Bocchieri, 1979); the chief difference, aside from the weather elements involved, is his evaluation of a later observation used directly and blended with MOS, and ours of GEM (based on a later observation) blended with MOS.

To blend the two systems, we derived eight multiple regression equations. These equations represent the elements of ceiling and visibility, each for two forecast projections (3 and 9 hours). All of these equations were derived to provide guidance under two situations: when current cycle MOS guidance is available, and when only previous cycle MOS guidance is available. The equations are of the REEP (Regression Estimation of Event Probabilities, see Miller, 1964) form and use as predictors five of the six GEM and MOS probabilities. The probability of one category is omitted because of redundancy, since each of the forecast processes (GEM and MOS) sum to one.

Table 30 presents the coefficients and additive constants of the REEP blending equations for the element ceiling for 3- and 9-h projections. The visibility equation elements have a similar form.

The GEM probabilities derived from GEM 03 GMT input data were paired with the MOS probabilities derived from input data from the previous MOS cycle (using previous 15 GMT observational predictors and previous 12 GMT model predictors), while GEM probabilities derived from GEM 15 GMT input data were paired with MOS probabilities derived from the input data also from the previous MOS cycle (using previous 03 GMT observational predictors and previous 00 GMT model output as predictors). Data so paired were aggregated from both cycles to achieve the REEP blending equations of Table 30. The pairing of GEM and MOS probabilities just described corresponds to the special operaional comparison defined in Section 2 and used in the GEM-MOS comparative verification of Section 3.

*Present affiliation: National Weather Service Forecast Office, Washington, D.C.

Coefficients and additive constants of the REEP blending equations were also formed for ceiling, for 3- and 9-h projections, with pairings of GEM and MOS probabilities corresponding to the operational comparison defined in Section 2 and used in Section 3. The GEM probabilities were derived from GEM 09 GMT input data paired with MOS probabilities derived from input data of the same MOS cycle (03 GMT observational predictors and 00 GMT model predictors), while the GEM probabilities derived from GEM 21 GMT input data were paired with MOS probabilities derived from input data of the same MOS cycle (15 GMT observational predictors and 12 GMT model predictors). Data so paired were then aggregated from both cycles to produce the operational comparison REEP blending equations.

A similar set of coefficients and additive constants were also prepared for visibility (i.e., for special operational and operational comparisons for both 3- and 9-h projections).

The REEP blending equations were applied to the MOS and GEM probability data to produce new estimates of probability for each category of ceiling and visibility for 3- and 9-h projections, under the special operational and operational comparisons. These new probability forecasts were converted to categorical forecasts by applying a P-star thresholding process similar to that used in GEM (see Miller, 1981). Results of the blending forecast process are displayed in Tables 31 and 32.

In general, the blend of GEM and MOS is an improvement over either forecast process alone. This is true for both ceiling and visibility at 3- and 9-h projections, and under both the special operational and operational conditions, for all measures except the chi-square (see Tables 31 and 32). Of the measures other than the chi-square, (except for the threat scores under the special operational comparison at the 9-h projection for both ceiling and visibility, and the threat score under the operational comparison at the 9-h projection for ceiling), the percentage improvement of the blended forecast process is greater over MOS for the 3-h projection and greater over GEM for the 9-h projection. These improvements are accompanied in some instances by smaller chi-square values for the blended results in comparison to either MOS or GEM alone.

In two cases, the special operational situation and the operational situation for ceiling at 3 hours, the reduction in chi-square achieved by the blended process over GEM alone, was on the order of 100 (see Tables 31 and 32). Occasionally the blended forecast technique achieves larger (less desirable) chi-square values than for MOS and GEM alone. The largest unfavorable difference between the blended chi-square values and that from either GEM or MOS alone is only 10.76.

The results of blending suggest that:

a. GEM alone is slightly improved by the blending process at small projections, such as 3 hours.

b. MOS alone is significantly improved by the blending process at larger projections, such as 9 hours and beyond.

These improvements are accompanied in some instances by better forecast balance, and in some instances by worse balance, as indicated by the chi-square measures.

Some caution must be exercised in comparing the results of blending with either unblended GEM or MOS alone. The GEM and MOS results reported in Section 3 represent verification on independent data, while the blending experiment results reported in this section represent regression fits on that same "independent" data. It would be quite reasonable to expect some "shrinkage" of the blending experiment results if verified on real independent data. However, the amount of shrinkage is deemed to be minor since the number of fitted regression coefficients is very small compared to the sample sizes.

Blending of the sort described here would be an interesting medium for combining MOS and GEM. Blending is an implicit way of utilizing the crossover information reported in the previous section of this report to improve upon the results of either product alone. Also, blending obviates conflicts in the guidance offered by the two systems separately. Blended forecasts would be self-consistent from projection to projection.

Some difficulties, though, might arise in implementing blending. To blend MOS into GEM, yet preserve GEM's capability to forecast for any hour, would require separate blending equations for each difference between the time of the observation used as input to GEM and the time of the observational predictors used in MOS. This requirement follows because centralized MOS is fixed in time as it is generated only twice daily, while GEM is not. The amount of computer storage necessary to hold all the blending equations might be so large as to increase beyond acceptable limits GEM's size for mini-computer applications. Consider, though, blending GEM into MOS. If only the blending equations from this section's experiment were used, it would be possible to issue "updated" MOS guidance shortly after 03 GMT (15 GMT) (based on the previous cycle) and after 09 GMT (21 GMT) (based on the current cycle). Fresh guidance at these times appear to be important for aviation support, in view of FT file times (1540 GMT, 0940 GMT, and 2140 GMT in continental U.S. NWS locations).

5. FEEDBACK

Following completion of the GEM-MOS comparative verification, we examined closely the residuals of the GEM and MOS forecasting processes. The term "residual" is commonly used in meteorological statistics to refer to the difference between what was forecast (usually by a regression, or regression-like process, such as underlies both MOS and GEM) and what was actually observed. Contemplation of residuals, to gain insight into the performance of a regression fitting process, is strongly advocated by data analysts such as Tukey (1977) and experts in regression analysis such as Draper and Smith (1981).

We chose for detailed analysis the data set for the element temperature under the scientific comparison for the warm season/15 GMT GEM input time (12 GMT MOS cycle) stratification. Temperature was an element in which GEM performed less well than MOS. GEM results for the warm season/15 GMT stratification, while neither the best nor the worst in comparison to MOS, nonetheless are typical of the differences in performance between the forecast processes.

To view the geographical distribution of GEM's mean algebraic error, see Fig. 4. We hesitate to display the data in contoured form, since the errors displayed may be highly location-specific, and the values between the discrete points displayed may not be appropriately represented by contour intervals. In any event Fig. 4 indicates that for the 21 stations in the data sample, there is considerable station-specific "bias" in the GEM temperature forecasts.

An obvious first step in employing feedback is simply to correct each station's temperature forecasts by the station's mean algebraic error. This assures that the station's mean algebraic error for the sample is reduced to zero, and likely also reduces somewhat the mean absolute error, the goal of applying feedback. Before applying this first step, however, we looked ahead to experiments with other, more complex forms of feedback, and selected a subsample of the temperature data which includes only those data samples for which GEM temperature forecasts and verifying observations from the previous day were also available. Column one of Table 33 lists the 21 stations by call letter; column two gives the selected subsample size; column three displays, as a reference, the mean absolute error achieved by unadorned GEM*; and column four indicates the mean absolute errors achieved by feeding back, for each station, its mean algebraic error. The weighted average of the mean absolute error for all 21 stations achieved by unadorned GEM is displayed in the final row of column three and that achieved by feeding back the station "bias" is displayed in the final row of column four. We were encouraged by an overall reduction in the weighted average of the mean absolute error of 0.40°F through application of this simple feedback scheme. The improvement is 9.2% over unadorned GEM.

Seeking to achieve greater reduction in the mean absolute error, we tried three sophisticated, but conceptually related, feedback schemes. In these applications of the feedback principle, we used feedback information contained in the error from the previous day's GEM forecasts. All three of these feedback schemes are of the general form:

feedback correction = $A + B\varepsilon$,

where ε is the previous day's GEM temperature forecast error, that is, the difference between the previous day's GEM temperature forecast and the actual observed temperature for the verifying hour; and A and B are regression coefficients.

In the first of the three error feedback schemes, the values of A and B (-0.13 and 0.41, respectively) are determined from data for the 21 stations taken together. The mean absolute temperature error results for this scheme are given in column five of Table 33. In the second error feedback scheme, the value of B is the same for all stations (0.33), but the value of A varies for each station. The resulting mean absolute error is given in column six Table 33. In the third error feedback scheme, the values of A and B are

*See Appendix A, Definition of Terms under "unadorned GEM."

determined separately for each station solely from its own data. These mean absolute error results are shown in column seven of Table 33.

Fig. 5 (see Miller, 1981, page 50) demonstrates in graphical form some of the similarities and differences among the three error feedback schemes. Fig. 5a shows the form of the first error feedback scheme, in which all the data from the 21 stations are grouped together and a single regression line has been fitted to these data. Fig. 5b shows the form of the second error feedback scheme, in which a single slope is derived for all stations from all station's data taken together, but a separate intercept is determined for each station. Fig. 5c shows the form of the third error feedback scheme, in which both slope and intercept have been individually fitted to each station's data.

The overall results, summarized by the weighted average of the mean absolute error for columns five, six, and seven of Table 33, indicate decreasing mean absolute error as the regression method becomes more station-specific. The best overall reduction in GEM mean absolute error (column seven of Table 33 compared with column three) is 0.76°F, an improvement of 17.5% over unadorned GEM.

Comparatively for MOS, mean absolute errors (station-by-station and weighted average), for MOS without feedback are displayed in column eight of Table 33. The mean absolute temperature error results obtained by applying to MOS the most station-specific of the three error feedback schemes (as in Fig. 5c) are given in column nine of Table 33. The weighted average reduction in the MOS mean absolute temperature achieved by applying this third error feedback scheme is 0.16°F, a 5.3% improvement.

The results in Table 33 should be viewed with some caution, however. The unadorned GEM and MOS error statistics are derived from independent test data, while the statistics documenting the application of the feedback processes result from dependent data. Consequently, we expect some shrinkage in the improvements resulting from application of feedback to independent data.

The benefits to be obtained from applying feedback are accompanied by some costs. The smallest improvements, resulting from simple feedback of the mean algebraic error, cost the least. To obtain the mean algebraic error, a representative sample of GEM forecasts and verifying observations is needed. The more sophisticated schemes, which employ feedback of the previous day's forecast errors, perform better, but at higher cost. With these schemes, not only must a properly constructed sample be used to derive regression coefficients, but the previous day's GEM forecast and verifying observation must also be available and carried along by the GEM forecast process. The need to carry along this additional information complicates somewhat GEM's straightforwardness as a forecasting procedure.

There appears, however, to be substantial benefit available, at little real additional cost, when error feedback is used with MOS: The MOS temperature improvement is on the order of 5%. The cost of carrying along the previous day's forecast and verification temperature is relatively small when viewed from the perspective of the large, centralized computer environment which produces MOS.

6. MISCELLANEOUS

This section describes some additional methods employed to attempt to improve GEM's probabilistic and categorical forecasting capabilities. Discussed are: (a) a collection of multivariate statistical techniques, related to one another, which we tested in making categorical and probabilistic forecasts; (b) an attempt to account more fully for local station effects by blending local monthly climatology along with local hourly climatology as part of the GEM probabilistic forecasting procedure; and (c) a very promising variation on GEM's P-star thresholding process, which we term unaccumulated P-stars.

A. Multivariate Statistical Forecasting Techniques

These multivariate statistical classification techniques have in common the need for mean matrices, derived for each forecast projection, and a variance-covariance matrix. As described in Miller (1981), the following relationship holds for a one-hour projection:

$$(\underline{Z}'\underline{Z})\ \underline{A} = (\underline{Z}'\underline{Y})_1,$$

where $\underline{A}$ is the GEM equation matrix, $(\underline{Z}'\underline{Z})$ is the predictor cross-product matrix and $(\underline{Y}'\underline{Z})$ is the predictand-predictor cross-product matrix. For a T hour projection, the following approximation is assumed to hold:

$$(\underline{Z}'\underline{Z})\ A^T \cong (\underline{Z}'\underline{Y})_T.$$

The required mean matrices for the Tth projection are derived from the $(\underline{Z}'\underline{Y})_T$ matrix (precise for the one hour projection, approximate for other projections). The required variance-covariance matrix is derived from $(\underline{Z}'\underline{Z})$. The mean vector for a particular element has as many rows as the number of GEM categories for that element. Each vector, corresponding to a GEM forecast category for that element, contains the mean value of each predictor, conditioned on the occurrence of that category.

The two quantities, the mean vector $(\underline{\mu})$ and the variance-covariance matrix $(\underline{\Sigma})$ are at the heart of each of the related categorical forecasting schemes with which we experimented. These quantities are an intergral part of the multivariate normal probability density function (See Miller, 1962, pp 6-9 for more details). We used this multivariate-normal approach to make categorical forecasts for certain elements using the probabilities produced by the GEM forecast process as input. This approach is parsimonious, in that it reduces the multivariate dimensionality involved from 228 (the number of binary predictands in the GEM process) to the number of categories for the element being predicted (at most 21, for wind*). In this new 21 dimensional space the non-orthogonal axes are the probabilities of the events as forecast by GEM.

One categorical decision-making scheme uses the concept of geometric distance within the element's probability space. For a given element, say

*Temperature, the element with the largest number of categories, 30, was not considered a candidate for this type of categorical decision process.

wind, with 21 categories, the element's categorical space contains 21 dimensions, and each of the 21 elements has a point (a centroid) in the 21 dimensonal space.

A GEM wind probability forecast is a single point in the 21 vector space, determined by the forecasted GEM probabilities for each of the 21 categories. The geometric (Euclidian) scheme calculates, in a straightforward geometric sense, the distance between the GEM forecasted point and each of the 21 centroids, and assigns the forecast to the category whose centroid is "closest", in the Euclidian sense, to the point represented by the GEM probability forecast.

Another categorical decision-making scheme employs a refinement of the Euclidian distance concept, using $\underline{\Sigma}$, called Mahalanobis distance. A spin-off from the Mahalanobis-distance scheme of categorical decision-making is a refinement of the probabilities of each element's categories. The refined, or a posteriori probability, is defined as the probability given that the forecast process was employed, and is obtainable from the Mahalanobis distance (for more details, see Miller, 1962, pp 6-9). Multivariate statistical theory suggests that such a posteriori probabilities should be "sharper" (i.e. should produce lower Brier Scores) than the GEM-forecasted probabilities used as input into the process, when there is underlying multivariate normality in the distribution.

Neither the Euclidian nor Mahalanobis distance classification schemes resulted in better categorical forecasts than the extant GEM P-star thresholding process, for the elements of wind or total cloud amount, suggesting a lack of multivariate normality. Also, the a posteriori probabilities resulting from the Mahalanobis-distance procedure were not, as measured by Brier Scores, "sharper" than the input GEM-forecasted probabilities.

We tried weighting the GEM forecast probabilities and the Mahalanobis-distance process a posteriori probabilities together using the weightings in Table 34.

None of the weightings produced either better Brier Scores nor categorically, a larger number of correct forecasts, than achieved by using GEM alone.

We remain optimistic, though, that some improvement in making categorical forecasts may follow from application of multivariate statistical principles, and we continue our search for and evaluation of these kinds of categorical decision-making techniques. We feel that GEM's respectable forecasting ability, reflected in its Brier Scores, suggest a corresponding potential for categorical forecasting improvement.

B. Station-Specific Monthly Climatological Corrections

To more fully account in the GEM forecast process for local station effects, we tried to blend together adjustments for local monthly climatology with adjustments for local hourly climatology. GEM, as comparatively verified against MOS in this study (see section 3), contains an adjustment to the forecast probabilities for the effects of station-specific (local) hourly climatology (for more details see Miller, 1981, pg 77-79). In particular, examination of the "residuals"[1] for the element wind revealed a strong bias in the wind forecasts for Los Angeles International Airport (LAX) that were seasonal in nature. We reasoned that if we could somehow adjust the GEM forecast probabilities on a station-specific basis for the month of the year, both the Brier Score and the GEM categorical forecasts should improve. After deriving the station-specific monthly climatological corrections for all 21 stations in the GEM-MOS comparative verification, we tried combining together the station- specific local and monthly climatological corrections into a single correction to the GEM forecast probabilities. Table 35 shows the weightings we used to achieve the combined corrections.

None of these weightings produced either better Brier Scores, or, categorically, a larger number of correct forecasts, than by using the station-specific hourly climatological corrections alone.

C. Unaccumulated P-Stars

The path that led to using unaccumulated P-stars in a categorical forecasting method began with the realization that use of accumulated P-stars[2] as categorical decision-making thresholds for the element wind, produced rather unattractive categorical forecasts. In the verification against MOS reported in Section 3 of this study, a maximum probability decision criterion was used to convert the GEM probabilistic forecasts to categorical ones. The maximum probability decision rule is not fully satisfying, however, because of the large chi-square measures associated with it (see part F, Section 3 of this study).

Examination of residuals revealed some evidence of seasonal effects at some stations, particularly Los Angeles International Airport (LAX), which led us in turn to look at the station-specific monthly climatological adjustments to the GEM probabilities reported in the preceding part of this section. Also tried, in conjunction with the element wind, were the multivariate statistical technqiues reported in the first part of this section. Neither of these approaches bore fruit.

One reason for our difficulties in categorical wind forecasts with GEM lies in our definition of the wind categories (see Section 2). These categories are a combination of wind speed and wind direction, and lack a unique ordering. A cumulative P-star method works best when the weather element possesses ordered categories.

An unaccumulated P-star represents a threshold probability for a single category, which if equalled by the GEM forecasted probability for that category, indicates that the category is likely to occur at a rate equal to its climatology in a suitable large sample. Accordingly, one requirement for a forecast decision rule might be that a category be selected if its

1 How well GEM categorical forecasts performed compared to verifying observations.

2 See Miller, 1981, Appendix for discussion of the accumulated P-star process.

forecasted probability exceeds its unaccumulated P-star threshold probability. More than one category may have its forecast probability exceed its P-star, however. A logical extension of the decision rule, then, is to pick the category whose forecast probability most exceeds the category's unaccumulated P-star threshold probability. And, in the event no single category's forecast probability exceeds the category's threshold unaccumulated P-star, pick the category whose probability lies closest to its threshold P-star.

Restated, the unaccumulated P-star process decision rule is:

a. Pick the category whose forecasted probability exceeds its P-star by the largest amount.

b. If no category's forecasted probability exceeds its P-star, pick the category whose forecasted probability is closest to its P-star.

When applied to the data samples used in the GEM-MOS verification, here are the results:

For 3-h projections, whether under the operational or scientific comparison, for all measures (except the Heidke Skill Score and percent correct for the warm season/21 GMT GEM input time stratification, under the operational comparison), the unaccumulated P-star method ("new" method) is better than maximum probability method ("old" method) (see Table 36).

For 9-h projections, the results are mixed, but for every stratification except warm season/09 GMT GEM input time under the operational comparision, the unaccumulated P-star method achieves a lower chi-square value than the maximum probability method, indicating better balance (see Table 37).

The use of unaccumulated P-stars, however, does not change the relative rankings of the forecasting performance between GEM and MOS. Use of the unaccumulated P-star, however, does reduce the chi-square values of GEM when compared with MOS, in some of the stratifications, by rather large amounts.

7. CONCLUSIONS

GEM demonstrates improvements in forecasting skill over MOS, particularly under the special operational and operational comparisons used in this study. This improvement is strongest for the elements most crucial for aviation operational forecasting (major aviation elements): ceiling, visibility, and total cloud amount. GEM's improvement over MOS, though present, is somewhat less pronounced for the remaining lesser aviation elements considered in this study: temperature, dewpoint depression, and wind. As indicated by the crossover tables of Section 3 of this study, for the major aviation elements, the crossover, determined within the resolution of the data used, lies between 5 and 8 hours from the time of the reference input observation. For the lesser aviation elements, the crossover lies between 3 and 5 hours. Based on the results of this study and previous comparisons of GEM with persistence, we conclude that GEM posseses considerable skill of value for short range operational forecasting guidance.

8. ACKNOWLEDGEMENTS

We wish to thank Albert Forst for his valuable help in preparing MOS forecasts and for securing observational data necessary for making GEM forecasts. Advice and assistence came also from Karl Hebenstreit, Bill Herrmann, George Hollenbaugh, and Bob Rood. Typing was performed by both Mary Battle and Anne Falkenhof to whom we wish to express our thanks.

9. REFERENCES

Bocchieri, J. R., 1979: Experiments in the use of local surface and upper-air observations to update MOS precipitation type guidance. TDL Office Note 79-14, National Weather Service, NOAA, U.S. Department of Commerce, 15 pp.

_____, 1982: Personal communication.

Carter, G. M., J. R. Bocchieri, and J. P. Dallavalle, 1982: Applications of the MOS technique: A bibliography - No. 1. TDL Office Note 82-10, National Weather Service, NOAA, U.S. Department of Commerce, 33 pp.

Draper, N. R., and H. Smith, 1966: Applied Regression Analysis. John Wiley and Sons, New York, 407 pp.

Feller, W., 1950: An Introduction to Probability Theory and Its Applications. John Wiley and Sons, New York, 419 pp.

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.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. Meteorological Monographs, Amer. Meteor. Soc., Vol. 4, No. 25, 54 pp.

_____, 1964: Regression estimation of event probabilities, Technical Report, The Travelers Research Center, Inc., Contract No. Cwb 10704, Hartford, Conn., 153 pp.

_____, 1968: A stochastic model for real-time on-demand weather prediction. Proceedings First Statistical Meteorological Conference, Hartford, Amer. Meteor. Soc., 48-51.

_____, 1979: Conditional climatology given the local observation. Nat. Wea. Digest, 4, 2-15.

_____, 1981: GEM: A statistical weather forecasting procedure. NOAA Techical Report NWS-28, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 103 pp.

_____, T. J. Perrone, and W. C. Herrmann, 1981: Details on the application of GEM. Preprints Seventh Conference on Probability and Statistics in Atmospheric Sciences, Monterey, Amer. Meteor. Soc., 166-169.

_____, R. C. Whiton, and M. J. Kelly, Jr., 1977: Results of a single station forecasting experiment. Proceedings Fifth Conference on Probability and Statistics in Atmospheric Sciences, Las Vegas, Amer. Meteor. Soc., 37-40.

Tukey, J. W., 1977: Exploratory Data Analysis. Addison-Wesley Publishing Co., Reading, Mass, 688 pp.

Whiton, R. C., 1977: Markov processes. Selected Topics on Statistical Meteorology, Ed., R. G. Miller, Air Weather Service Tech. Report AWS-TR-77-273, Scott AFB, Ill., 45 pp. (Available from Air Weather Service Headquarters, Scott AFB, Ill.)

Wiener, N., 1948: Cybernetics. The Technology Press, MIT, John Wiley and Sons, New York, 194 pp.

_____, 1950: Extrapolation, Interpolation, and Smoothing of Stationary Time Series. The Technology Press, MIT, John Wiley and Sons, New York, 163 pp.

_____, 1956: Nonlinear prediction and dynamics. Proceedings of the Third Berkeley Symposium, Ed., J. Neyman, University of California Press, Berkeley, 247-252.

Appendix A
Definitions of Terms

AFOS: Automation of Field Operations and Services. NWS field station computer system which stores and displays centrally-prepared forecast products as well as collectives of weather observations. Allows field forecasters to compose forecast and warning text products and transmit them to users, and posseses limited capability to run on-site applications programs.

AWS: Air Weather Service. Weather forecasting agency of U.S. Air Force.

Backup Equations: MOS forecast equations derived solely with model output parameters as predictors. Used operationally when surface observations are unavailable.

Bias: Equal to the number of forecasts of an event divided by the number of times the event occurred. A bias of 1.0 is perfect, less than one implies underforecasting, greater than one, overforecasting.

Binary Variable: A variable having a value zero or one. Binary variables, such as used in GEM, are also called "dummy" variables.

Blending: A regression technique which uses predicted probabilities produced by both GEM and MOS for all (less one) categories of ceiling as predictors to produce a new refined probability forecast for each ceiling category. A similar procedure is used for visibility.

Brier Score: $[\Sigma_{\text{all events}}(\text{Probability of an event} - (\text{one, if event occured, zero, if it did not}))^2]/(2\cdot \text{number of cases})$. Lower values are preferred.

Chi Square Goodness-of-Fit: Tests goodness-of-fit between observed and forecast frequencies. Since both frequencies are only estimates of true frequency of occurrence, this test uses expected values based on the average of observed and forecast frequencies. Chi-Square = Σ[(observed frequency - expected frequency)^2/(expected frequency)]. A smaller score is to be preferred.

Classical statistical approach: Statistical weather prediction using observational elements as predictors and predictands. No model output is used as a source of predictors.

Cool Season: October 1st to March 31st. The coolest one-half of the year.

Cost Benefit Analysis: A quantitative management science technique which weighs costs against benefits. Used here in a qualitative way to identify and to compare costs with benefits.

Crossover point: The forecast projection where GEM and MOS forecast guidance are of equivalent value. For projections less than the crossover point, GEM is superior to MOS; for projections greater, the opposite is true.

Cycle: The National Meteorological Center's numerical models run twice daily, once from data input to the models at 00 GMT and again at 12 GMT. MOS is produced with output from each model run, hence the terms 00 GMT MOS cycle and 12 GMT MOS cycle.

Feedback: A technique developed in electrical engineering, later popularized by Wiener and broadened to other fields, which uses error information to control, limit and/or improve a process. Here, temperature forecast error information was used to limit future errors in GEM and MOS temperature forecasts.

Fraction: A way of expressing relative skill between GEM and MOS for a number of scores under the scientific, operational, and special operational comparisons. The numerator is the number of scores for which GEM is superior to MOS; the denominator is the number of non-chi-square scores for a weather element.

FT: Aviation terminal forecast. The letters "FT" are used in communications bulletin headings. A forecast for aviation operations which predicts elements such as ceiling, visibility, clouds, present weather and obstructions to vision, and wind.

GEM: Generalized Exponential Markov. A short-range, purely statistical weather forecasting technique, which requires only a surface observation and local climatology as input.

GEM input time: The time of an observation used as an input to GEM. Here, GEM input times of 03 GMT and 15 GMT are used for the operational and special operational comparison; 09 GMT and 21 GMT GEM input times are used for the scientific comparison.

Heidke Skill Score: $\Sigma_{\text{all categories}}$[(Hits - Expected Hits due to chance)/ (Total number of cases - Expected Hits due to chance)].

Hits: Number of correct forecasts.

LFM: Limited-Area Fine Mesh. A dynamic modeling system which, over the United States and nearby contiguous regions, uses a smaller grid-length ("mesh") than the National Meteorological Center's hemispheric and global models. Some predictors from the LFM are used in MOS.

Mean absolute error: $\Sigma_{\text{all cases}}$[absolute value (Forecast - Observed)]/Number of cases.

Mean algebraic error: $\Sigma_{\text{all cases}}$[(Forecast-observed)]/Number of cases.

MOS: Model Output Statistics. A dynamical-statistical weather forecasting technique which uses model output and surface and upper air observations as predictors.

Number of Large Errors: A count of events where the forecast and observed temperatures differ by 10°F or more.

OBS: Surface Weather Observations.

Operational Comparison: Method of comparing MOS and GEM where different observational predictors are used by GEM and MOS. GEM uses later observational information than MOS, at a time chosen to be approximately one hour before FT file time. For example, MOS uses 03 GMT surface observational parameters while GEM uses those at 09 GMT.

P-star (P*): A probability value, which, if exceeded by the forecast probability, would initiate a categorical forecast of the event.

Percent correctly forecast: The number of "Hits" divided by the total number of cases, expressed as a percentage.

REEP: Regression Estimation of Event Probabilities. A regression technique where the predictands are only zero or one. The objective is to estimate the probability that the event one will occur.

Residuals: The difference between the fit produced by regression on data, and the data values themselves. Analysis of residuals can provide indications of distributional forms and biases in the regression analysis, and can suggest ways to improve regression fit.

Scientific comparison: Method for comparing MOS and GEM where both forecast techniques use the same surface observational parameters as predictors.

Shrinkage: Degradation of forecast performance on independent data when compared with performance on dependent data. Shrinkage is small when the number of cases in the data sets is large compared with the number of terms fitted by the regression.

Special operational comparison: Method for comparing MOS and GEM where different observational predictors are used by GEM and MOS. MOS uses data from the previous cycle's model run. This comparison simulates the situation in field forecast offices, where current cycle MOS guidance is not yet available. For example, GEM uses 03 GMT observational parameters, while MOS uses the previous cycle's 12 GMT model input and 15 GMT observational parameters.

Threat Score: Hits/(Number of forecasts + number of observed cases -Hits).

Unaccumulated P-star: A categorical decision-making procedure for unordered events.

Unadorned GEM: GEM without blending, feedback, or unaccumulated P-star enhancements (See Miller (1981), Section 7, pp 73-79).

Warm season: April 1st to September 30th, the warmest one-half of the year.

Weighted average: The sum of: each value of a categorical element, such as temperature, multiplied by its GEM probability estimate of occurrence.

Table 1. Ceiling category definitions.

Category Number	Category Definition (ft)
1	<200
2	200-400
3	500-900
4	1000-2900
5	3000-7500
6	>7500

Table 2. Visibility category definitions.

Category Number	Category Definition (mi)
1	<1/2
2	1/2-7/8
3	1-2 1/2
4	3-4
5	5-6
6	>6

Table 3. Temperature category definitions.

Category Number	Category Definition (°F)
1	-130 - -31
2	-30 - -26
3	-25 - -21
4	-20 - -16
5	-15 - -11
6	-10 - - 6
7	-5 - - 1
8	0 - 4
9	5 - 9
10	10 - 14
11	15 - 19
12	20 - 24
13	25 - 29
14	30 - 34
15	35 - 39
16	40 - 44
17	45 - 49
18	50 - 54
19	55 - 59
20	60 - 64
21	65 - 69
22	70 - 74
23	75 - 79
24	80 - 84
25	85 - 89
26	90 - 94
27	95 - 99
28	100 - 104
29	105 - 109
30	110 - 140

Table 4. Dewpoint depression category definitions.

Category Number	Category Definition (°F)
1	0
2	1
3	2 - 4
4	5 - 7
5	8 - 11
6	12 - 15
7	16 - 19
8	20 - 25
9	26 - 35
10	36 - 50
11	51 - 99

Table 5. Wind category definitions.

Category	Wind Direction (°) and Speed (kt)
1	Calm or less than 2
2	020 - 050/2-9
3	020 - 050/10-19
4	060 - 100/2-9
5	060 - 100/10-19
6	110 - 140/2-9
7	110 - 140/10-19
8	150 - 190/2-9
9	150 - 190/10-19
10	200 - 230/2-9
11	200 - 230/10-19
12	240 - 280/2-9
13	240 - 280/10-19
14	290 - 320/2-9
15	290 - 320/10-19
16	330 - 010*/2-9
17	330 - 010*/10-19
18	020 - 100/>20
19	110 - 190/≥20
20	200 - 280/≥20
21	290 - 010*/≥20

* through 360

Table 6. Station list for GEM-MOS comparative verification.

Station Name	Station Identifier
Washington, D.C. (Nat. Airport)	DCA
Boston, Mass.	BOS
Buffalo, N.Y.	BUF
Philadelphia, Pa.	PHL
Raleigh-Durham, N.C.	RDU
Knoxville, Tenn.	TYS
Savannah, Ga.	SAV
Miami, Fla.	MIA
New Orleans, La.	MSY
St. Louis, Mo.	STL
Milwaukee, Wis.	MKE
Minneapolis-St. Paul, Minn.	MSP
Bismarck, N.D.	BIS
Oklahoma City, Okla.	OKC
San Antonio, Tex.	SAT
Denver, Colo.	DEN
Albequerque, N.M.	ABQ
Salt Lake City, Utah	SLC
Portland, Oreg.	PDX
San Francisco, Calif.	SFO
Los Angeles, Calif.	LAX

Table 7. Ceiling comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Ceiling

Operational Comparisons

Projection: 3 hours

	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.161	.139	.191	.153	.112	.102	.163	.136	.161	.135*
% Corr	73.1	80.4	68.6	78.6	80.8	85.4	73.6	79.9	73.4	80.7*
Heidke	.380	.487	.394	.556	.391	.419	.418	.500	.396	.497*
Chi Sq	1.98	29.6	3.61	16.9	2.81	30.3	2.82	54.2		
Threat	.184	.271	.241	.404	.120	.250	.140	.321	.201	.345*
Sample Size	2720		3194		2141		3091		11146	

Scientific Comparisons

Projection: 3 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.108	.108	.142	.142	.128	.133	.158	.161	.136*	.138
% Corr	82.4	85.1	78.3	79.5	81.0	79.8	76.5	75.3	79.3	79.7*
Heidke	.494	.472	.538	.513	.532	.413	.537	.477	.526*	.472
Chi Sq	4.52	25.7	.79	29.9	1.28	50.1	4.36	55.7		
Threat	.350	.360	.366	.354	.208	.159	.362	.322	.351*	.319
Sample Size	2518		3187		2435		3116		11256	

* Signifies superiority (shown only for aggregated)

Table 8. Ceiling comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 9 hour after GEM input time.

Element: Ceiling

Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.164	.175	.190	.201	.135	.143	.181	.191	.171*	.181
% Corr.	74.3	72.0	68.9	65.6	76.8	79.3	70.3	69.2	72.1*	70.8
Heidke	.347	.281	.373	.341	.356	.313	.382	.348	.366*	.288
Chi Sq.	4.70	13.8	8.51	18.1	2.54	23.2	4.02	8.43		
Threat	.057	.159	.117	.162	.042	.077	.163	.212	.121	.158*
Sample Size	2726		3198		2139		3091		11154	

Scientific Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.162	.180	.192	.210	.110	.122	.164	.184	.160*	.177
% Corr.	72.7	73.8	68.5	63.5	81.3	80.1	73.4	67.8	73.6*	70.6
Heidke	.377	.346	.393	.319	.391	.281	.421	.311	.397*	.315
Chi Sq.	2.22	25.2	4.42	6.70	3.99	3.75	5.72	25.4		
Threat	.180	.118	.243	.209	.038	.000	.134	.147	.195*	.166
Sample Size	2524		3185		2428		3107		11244	

*Signifies superiority

Table 9. Ceiling comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Element: Ceiling

Special Operational Comparisons

Projection: 3 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.127	.107	.179	.144	.155	.133	.193	.161	.166	.138*
% Corr.	78.4	85.3	70.8	79.3	76.3	79.8	68.5	75.2	73.0	79.6*
Heidke	.368	.463	.382	.511	.381	.413	.374	.474	.377	.469*
Chi Sq.	2.60	27.5	.97	29.6	2.72	48.1	8.48	56.9		
Threat	.060	.386	.174	.362	.077	.182	.111	.319	.126	.325*
Sample Size	2458		3113		2438		3190		11199	

Special Operational Comparisons

Projection: 9 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.167	.179	.205	.212	.115	.124	.172	.185	.168*	.178
% Corr.	72.6	73.8	65.8	63.2	80.1	80.1	72.2	67.7	72.2*	70.5
Heidke	.371	.340	.345	.317	.341	.290	.372	.309	.358*	.314
Chi Sq.	2.45	28.1	7.86	6.25	17.5	4.44	6.42	28.5		
Threat	.132	.122	.170	.219	.024	.000	.096	.143	.137	.169*
Sample Size	2460		3110		2429		3179		11178	

*Signifies superiority

Table 10. Ceiling contingency tables for the operational comparison of GEM and MOS. GEM forecasts use 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Elememt: Ceiling
Projection: 3 hours
Operational Comparison

	MOS Forecasts						GEM Forecasts						
CATS.	1	2	3	4	5	6	1	2	3	4	5	6	TOTAL
1	12	22	7	4	5	26	23	24	6	0	1	22	76
O 2	32	43	46	36	26	51	9	85	54	26	9	51	234
B 3	14	77	100	143	40	70	7	38	183	115	18	83	444
S 4	10	48	109	367	214	233	5	21	75	482	109	289	981
5	4	16	35	205	459	532	0	6	18	104	402	721	1251
6	18	46	73	250	569	7204	2	20	29	97	188	7824	8160
TOTAL	90	252	370	1005	1313	8116	46	194	365	824	727	8990	11146
BIAS	1.18	1.08	.83	1.02	1.34	.99	.61	.83	.82	.84	.58	1.10	

MOS:
Brier Score = .161
% Corr = 73.4
Heidke = .396
Threat Score = .201
Sample Size = 11146

GEM:
Brier Score = .135
% Corr = 80.7
Heidke = .497
Threat Score = .345
Sample Size = 11146

Element: Ceiling
Projection: 9 Hours
Operational Comparison

	MOS Forecasts						GEM Forecasts						
CATS.	1	2	3	4	5	6	1	2	3	4	5	6	TOTAL
1	4	8	7	7	0	12	4	10	10	5	0	9	38
O 2	3	33	32	59	26	35	17	34	39	34	33	31	188
B 3	4	51	81	135	44	82	10	49	68	138	46	86	397
S 4	9	34	82	403	263	307	9	43	104	372	233	337	1098
5	1	25	25	226	429	612	7	12	38	195	345	721	1318
6	15	32	71	304	599	7094	17	39	76	313	595	7075	8115
TOTAL	36	183	298	1134	1361	8142	64	187	335	1057	1252	8259	11154
BIAS	.95	.97	.75	1.03	1.03	1.00	1.68	.99	.84	.96	.95	1.02	

MOS:
Brier Score = .171
% Corr = 72.1
Heidke = .366
Threat Score = .121
Sample Size = 11154

GEM:
Brier Score = .181
% Corr = 70.8
Heidke = .288
Threat Score = .158
Sample Size = 11154

Table 11. Visibility comparative GEM-MOS verification scores. Under the operational comparison GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations. Forecasts are valid 3 hours after GEM input time.

Element: Visibility

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.171	.171	.154	.132	.092	.077	.110	.090	.134	.119*
% Corr.	72.3	76.7	74.4	82.5	85.7	88.6	81.8	86.9	78.1	83.5*
Heidke	.342	.374	.301	.492	.368	.469	.321	.520	.330	.467*
Chi Sq.	6.63	55.3	10.4	14.6	1.49	5.86	3.78	1.01		
Threat	.129	.210	.239	.343	.000	.143	.211	.393	.202	.319*
Sample Size	2753		3208		2153		3129		11243	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.073	.073	.087	.086	.084	.091	.115	.118	.091*	.093
% Corr.	89.1	90.6	86.2	88.5	87.8	86.7	80.6	80.1	85.6	86.3*
Heidke	.494	.523	.459	.521	.521	.499	.429	.447	.472	.496*
Chi Sq.	.49	3.75	4.31	8.34	5.12	9.02	13.7	14.0		
Threat	.313	.278	.413	.474	.000	.000	.372	.324	.374*	.347
Sample Size	2553		3204		2443		3142		11342	

*Signifies superiority

Table 12. Visibility comparative GEM-MOS verification scores. Under the operational comparison GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations. Forecasts are valid 9 hours after GEM input time.

Element: Visibility

Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.106	.118	.134	.144	.099	.105	.113	.116	.114*	.122
% Corr.	83.3	79.9	79.1	75.5	83.5	85.7	80.3	80.4	81.3*	79.9
Heidke	.279	.173	.352	.271	.260	.198	.297	.290	.301*	.238
Chi Sq.	4.27	45.2	6.48	25.8	4.80	39.0	2.41	7.96		
Threat	.000	.000	.182	.173	.059	.000	.198	.220	.175*	.167
Sample Size	2753		3208		2153		3129		11243	

Scientific Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.178	.208	.154	.176	.090	.101	.110	.118	.134*	.151
% Corr.	71.6	64.4	74.3	62.9	85.8	85.8	81.7	78.2	78.2*	72.4
Heidke	.342	.218	.302	.171	.355	.187	.322	.262	.328*	.210
Chi Sq.	7.59	47.8	12.2	48.9	2.21	32.4	3.70	16.7		
Threat	.126	.107	.196	.170	.000	.000	.198	.157	.177*	.152
Sample Size	2552		3204		2443		3142		11341	

*Signifies superiority

Table 13. Visibility comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after input time for 3-h projection; 9 hours, for 9-h projection.

Element: Visibility

Special Operational Comparisons

Projection: 3 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.091	.072	.109	.087	.112	.091	.135	.118	.113	.093*
% Corr.	84.4	90.7	81.0	88.3	82.4	86.7	78.7	80.1	81.4	86.2*
Heidke	.256	.530	.299	.518	.273	.500	.344	.443	.297	.496*
Chi Sq.	4.03	3.96	3.10	7.76	6.59	8.78	6.24	13.9		
Threat	.087	.333	.211	.481	.000	.000	.195	.313	.190	.348*
Sample Size	2472		3150		2465		3197		11284	

Special Operational Comparisons

Projection: 9 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.182	.203	.164	.176	.096	.099	.114	.117	.139*	.148
% Corr.	69.3	65.3	72.3	63.6	85.5	87.1	82.4	78.8	77.4*	73.4
Heidke	.321	.223	.274	.172	.247	.182	.269	.264	.277*	.211
Chi Sq.	1.29	46.5	9.89	48.2	4.91	31.6	11.3	16.9		
Threat	.099	.115	.169	.181	.000	.000	.060	.148	.131	.156*
Sample Size	2471		3150		2465		3197		11283	

*Signifies superiority

Table 14. Visibility contingency tables for the operational comparison of GEM and MOS. GEM forecasts use 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Element: Visibility
Projection: 3 Hours
Operational Comparison

		MOS Forecasts						GEM Forecasts						
CATS.		1	2	3	4	5	6	1	2	3	4	5	6	TOTAL
	1	24	20	18	8	10	29	47	9	19	6	8	20	109
O	2	25	11	22	12	9	21	15	16	39	9	4	17	100
B	3	44	27	129	98	113	131	21	21	192	148	73	87	542
S	4	16	20	98	107	127	228	5	6	77	182	119	207	596
	5	7	17	62	102	167	496	1	3	28	90	226	503	851
	6	25	32	107	198	336	8347	3	4	41	86	188	8723	9045
TOTAL		141	127	436	525	762	9252	92	59	396	521	618	9557	11243
BIAS		1.29	1.27	.80	.97	.90	1.02	.84	.59	.73	.87	.73	1.06	

MOS:
Brier Score = .134
% Corr = 78.1
Heidke = .330
Threat Score = .202
Sample Size = 11243

GEM:
Brier Score = .119
% Corr = 83.5
Heidke = .467
Threat Score = .319
Sample Size = 11243

Element: Visibility
Projection: 9 Hours
Operational Comparison

		MOS Forecasts						GEM Forecasts						
CATS.		1	2	3	4	5	6	1	2	3	4	5	6	TOTAL
	1	28	7	11	10	5	18	19	10	18	8	7	17	79
O	2	7	2	11	11	4	22	9	4	12	7	6	19	57
B	3	21	21	114	71	61	119	20	21	130	60	32	144	407
S	4	2	8	50	83	99	215	5	13	62	68	63	246	457
	5	2	9	42	58	152	478	7	0	52	66	71	545	741
	6	24	29	96	166	425	8762	25	25	126	340	294	8692	9502
TOTAL		84	76	324	399	746	9614	85	73	400	549	473	9663	11243
BIAS		1.06	1.33	.80	.87	1.01	1.01	1.08	1.28	.98	1.20	.64	1.02	

MOS:
Brier Score = .114
% Corr = 81.3
Heidke = .301
Threat Score = .175
Sample Size = 11243

GEM:
Brier Score = .122
% Corr = 79.9
Heidke = .238
Threat Score = .167
Sample Size = 11243

Table 15. Total cloud amount comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Total Cloud Amount

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.291	.291	.269	.237	.296	.272	.279	.240	.282	.258*
% Corr.	53.7	55.0	57.1	65.8	52.0	58.4	54.6	63.4	54.6	61.1*
Heidke	.367	.390	.407	.508	.337	.442	.388	.492	.378	.462*
Chi Sq.	1.05	78.0	42.8	80.3	5.74	7.99	68.1	4.81		
Sample Size	2735	2713	3208	3169	2153	2151	3129	3123	11225	11156

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.232	.246	.221	.225	.255	.289	.246	.252	.238*	.251
% Corr.	63.3	60.9	66.2	66.4	59.7	56.0	61.1	62.9	62.7*	62.0
Heidke	.468	.449	.514	.505	.445	.412	.482	.482	.480*	.466
Chi Sq.	7.60	6.89	19.2	59.0	20.6	49.6	57.2	13.8		
Sample Size	2536	2532	3204	3187	2443	2439	3142	3122	11325	11280

*Signifies superiority

Note: Sample sizes for GEM and MOS are slightly mismatched because of differences in total cloud amount definitions in the two forecast processes (see Section 2).

Table 16. Total cloud amount comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 9 hours after GEM input time.

Element: Total Cloud Amount

Operational Comparisons

Projection: 9 hours

	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.341	.341	.299	.321	.283	.313	.265	.288	.297*	.315
% Corr.	51.4	41.2	57.5	46.7	54.2	49.1	50.0	56.8	53.3*	48.6
Heidke	.315	.211	.337	.264	.345	.275	.387	.349	.348*	.274
Chi Sq.	5.5	18.1	62.8	41.3	2.81	11.0	30.5	121.		
Sample Size	2735	2729	3208	3186	2153	2150	3129	3114	11225	11179

Scientific Comparisons

Projection: 9 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.288	.329	.269	.297	.296	.340	.278	.312	.281*	.318
% Corr.	54.5	45.4	57.1	54.7	52.8	32.4	54.6	49.5	54.9*	48.4
Heidke	.380	.266	.407	.349	.343	.226	.387	.291	.382*	.287
Chi Sq.	.59	26.2	46.1	71.8	1.78	20.6	63.0	61.8		
Sample Size	2535	2512	3204	3165	2443	2440	3142	3135	11324	11252

*Signifies superiority

Table 17. Total cloud amount comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Element: Total Cloud Amount

Special Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.282	.247	.267	.225	.298	.288	.298	.254	.286	.252*
% Corr.	54.4	60.7	57.2	66.4	53.5	56.3	50.3	62.4	53.8	61.8*
Heidke	.343	.446	.384	.504	.342	.415	.340	.477	.354	.464*
Chi Sq.	5.12	6.41	24.9	54.4	3.62	45.8	62.1	14.3		
Sample Size	2472	2466	3150	3134	2446	2440	3197	3176	11265	11216

Special Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.304	.330	.286	.294	.308	.339	.294	.314	.296*	.318
% Corr.	51.2	45.1	52.5	54.9	49.2	41.5	51.4	48.9	51.2*	48.1
Heidke	.335	.263	.345	.349	.290	.220	.345	.284	.331*	.284
Chi Sq.	1.53	29.2	23.9	66.3	2.92	20.7	88.3	62.5		
Sample Size	2471	2451	3150	3112	2446	2443	3197	3190	11264	11196

*Signifies superiority

Table 18. Temperature comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

MABS Error = Mean Absolute Error
MALG Error = Mean Algebraic Error
No. LG Errors = Number of Large Errors (≥ 10°F)

Element: Temperature

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	2.12	2.10	2.90	2.23	3.03	2.80	2.92	2.97	2.73	2.51*
MALG Error	.20	-.30	.63	.07	-.32	-.86	-.15	.20	.33*	.36
No. LG Error	13	5	72	16	56	29	61	46	202	96*
% Correct	59.2	61.6	48.0	57.3	47.7	47.4	49.2	46.2	51.1	53.4*
Chi Sq.	7.61	11.0	12.6	5.37	6.02	20.6	6.71	12.4		
Heidke	.537	.564	.425	.528	.411	.407	.443	.411	.455	.481*
Sample Size	3031		3250		2380		3245		11906	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Projection Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	1.78	2.10	1.95	2.39	2.38	2.88	2.55	3.55	2.14*	2.75
MALG Error	.55	.26	.31	-.49	-.61	-.06	-.47	-1.2	.49*	.51
No. LG Errors	9	13	19	21	31	38	51	133	110*	205
% Correct	64.5	59.6	64.8	53.4	56.2	49.8	53.1	41.9	59.6*	50.9
Chi Sq.	5.92	13.3	5.34	9.94	12.8	27.8	8.76	30.6		
Heidke	.599	.541	.611	.486	.508	.435	.488	.363	.552*	.454
Sample Size	2788		3246		2691		3252		11977	

*Signifies superiority

Table 19. Temperature comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 6 hours after GEM input time.

MABS Error = Mean Absolute Error
MALG Error = Mean Algebraic Error
No. LG Errors = Number of Large Errors (≥ 10°F)

Element: Temperature

Operational Comparisons

Projection: 6 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	2.41	4.22	2.86	3.91	2.78	3.56	2.96	4.39	2.20*	4.05
MALG Error	.23	-1.92	.43	.90	-.55	-.03	-.08	-.32	.32*	.79
No. LG Errors	40	237	48	166	34	73	64	244	186*	720
% Correct	56.1	37.8	48.8	36.7	50.7	40.2	49.0	33.4	51.1*	36.8
Chi Sq.	12.4	44.4	12.1	27.7	9.98	36.5	8.56	27.7		
Heidke	.506	.302	.437	.302	.441	.319	.437	.267	.455*	.296
Sample Size	3031		3250		2381		3225		11887	

Scientific Comparisons

Projection: 6 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	1.97	2.95	2.56	3.44	3.06	4.30	3.11	4.85	2.68*	3.90
MALG Error	.33	.47	.43	-.99	-.53	-.20	-.41	-1.66	.43*	.83
No. LG Errors	9	42	42	112	73	199	104	408	228*	761
% Correct	61.4	46.6	54.3	42.7	47.6	35.2	46.4	31.8	52.3*	39.0
Chi Sq.	5.83	41.8	7.30	16.3	12.9	64.5	11.4	67.3		
Heidke	.563	.392	.495	.367	.412	.268	.415	.253	.470*	.260
Sample Size	2787		3246		2691		3252		11976	

*Signifies superiority

Table 20. Temperature comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 6 hours, for 6-h projection.

MABS Error = Mean Absolute Error
MALG Error = Mean Algebraic Error
No. LG Errors = Number of Large Errors ($\geq$ 10°F)

Element: Temperature

Special Operational Comparisons

Projection: 3 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	2.53	2.12	3.09	2.39	3.04	2.87	3.31	3.55	3.01	2.75*
MALG Error	-0.26	0.29	-0.02	-0.47	0.03	-0.09	0.08	-1.22	-0.02*	-0.41
No. LG Errors	22	12	87	21	89	39	103	134	295	206*
% Correct	52.7	59.5	46.7	53.3	48.6	50.1	44.7	41.8	48.0	50.9*
Chi Sq.	6.54	13.5	11.7	11.6	10.3	27.3	16.0	30.9		
Heidke	.463	.539	.411	.485	.423	.438	.395	.363	.421	.454*
Sample Size	2734		3244		2702		3241		11921	

Special Operational Comparisons

Projection: 6 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	2.49	2.96	3.27	3.45	3.48	4.30	3.52	4.85	3.21*	3.91
MALG Error	-0.19	-0.53	0.15	-0.96	-0.02	-0.26	-0.01	-1.67	-0.01*	-0.66
No. LG Errors	16	42	112	113	132	207	153	406	413*	768
% Correct	53.3	46.5	45.6	42.8	46.0	35.4	44.2	31.7	47.1*	38.9
Chi Sq.	4.91	40.8	14.4	16.0	11.9	65.5	19.3	68.7		
Heidke	.469	.389	.398	.368	.393	.270	.391	.252	.411*	.312
Sample Size	2733		3244		2702		3241		11920	

*Signifies superiority

Table 21. Dewpoint depression comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Dewpoint Depression

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	42.7	47.3	38.2	46.9	40.4	45.3	37.3	36.9	39.4	43.8*
Chi Sq.	57.1	68.7	103.	84.6	43.3	41.8	53.7	84.3		
Heidke	.272	.327	.227	.338	.312	.377	.272	.282	.267	.327*
Sample Size	2710		3250		2179		3245		11384	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	48.8	47.7	48.1	41.6	48.0	47.5	44.1	37.3	47.1*	43.0
Chi Sq.	18.4	26.6	26.9	42.8	30.7	68.6	50.2	169.		
Heidke	.378	.365	.373	.292	.393	.382	.357	.274	.374*	.322
Sample Size	2473		3246		2467		3252		11438	

*Signifies superiority

Table 22. Dewpoint depression comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 6 hours after GEM input time.

Element: Dewpoint Depression

Operational Comparisons

Projection: 6 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	40.7	30.7	34.9	29.8	35.6	30.1	34.4	23.9	36.2*	28.4
Chi Sq.	58.2	252.	72.7	402.	36.2	132.	81.4	235.		
Heidke	.295	.169	.231	.166	.238	.202	.212	.129	.242*	.163
Sample Size	2710		3250		2181		3225		11366	

Scientific Comparisons

Projection: 6 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	45.6	36.8	41.6	32.8	44.3	38.5	38.5	31.7	42.2*	34.6
Chi Sq.	38.6	70.7	76.0	63.5	45.1	119.	61.7	202.		
Heidke	.320	.220	.282	.176	.349	.271	.292	.205	.307*	.214
Sample Size	2473		3246		2466		3252		11437	

*Signifies superiority

Table 23. Dewpoint depression comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 6 hours for 6-h projection.

Element: Dewpoint Depression

Special Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	37.6	47.3	33.2	41.6	38.9	47.4	33.7	37.2	35.5	42.8*
Chi Sq.	43.4	26.0	111.	42.9	82.6	67.0	88.0	169.		
Heidke	.230	.361	.181	.293	.282	.381	.236	.273	.229	.321*
Sample Size	2495		3244		2446		3241		11426	

Special Operational Comparisons

Projection: 6 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	36.5	36.6	34.0	33.0	39.4	38.5	35.1	31.5	36.0*	34.5
Chi Sq.	75.0	73.6	145.	64.4	79.9	119.	101.	203.		
Heidke	.196	.219	.180	.177	.219	.272	.249	.203	.227*	.214
Sample Size	2495		3244		2446		3241		11426	

*Signifies superiority

Table 24. Wind comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Wind

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	35.0	38.4	32.7	39.7	33.7	29.6	33.6	27.6	33.7	34.0*
Chi Sq.	33.3	38.5	28.4	33.0	19.4	115.	44.1	255.		
Heidke	.282	.322	.268	.346	.287	.249	.284	.229	.279	.289*
Sample Size	3031		3250		2380		3245		11906	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	40.0	34.5	38.6	36.6	32.4	29.6	35.6	32.0	36.7*	33.3
Chi Sq.	34.5	91.3	24.4	66.1	16.9	34.5	23.5	50.7		
Heidke	.337	.285	.333	.316	.270	.242	.311	.274	.314*	.280
Sample Size	2788		3246		2691		3252		11977	

*Signifies superiority

Table 25. Wind comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 9 hours after GEM input time.

Element: Wind

Operational Comparisons

Projection: 9 hours

	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	29.3	19.5	30.0	17.9	33.4	23.2	31.5	18.8	30.9*	19.6
Chi Sq.	24.5	554.	24.6	792.	40.3	225.	36.9	335.		
Heidke	.237	.141	.252	.124	.265	.147	.255	.117	.252*	.131
Sample Size	3031		3250		2380		3226		11887	

Scientific Comparisons

Projection: 9 hours

	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	36.1	24.0	32.7	23.6	33.0	24.4	33.6	18.2	33.8*	22.4
Chi Sq.	26.9	141.	30.2	244.	26.4	430.	42.2	600.		
Heidke	.294	.160	.268	.167	.278	.178	.284	.123	.280*	.156
Sample Size	2787		3246		2691		3252		11976	

*Signifies superiority

Table 26. Wind comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Element: Wind

Special Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	33.1	34.6	31.4	36.7	29.7	29.1	29.9	31.9	31.0	33.2*
Chi Sq.	47.5	93.2	38.0	64.7	22.4	37.0	27.6	50.8		
Heidke	.261	.286	.253	.316	.241	.237	.251	.293	.252	.279*
Sample Size	2734		3245		2702		3241		11922	

Special Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	34.1	23.9	30.4	23.7	30.2	24.5	31.3	18.3	31.4*	22.5
Chi Sq.	27.4	138.	34.8	244.	23.3	435.	36.1	597.		
Heidke	.271	.158	.241	.168	.248	.178	.259	.124	.255*	.156
Sample Size	2733		3245		2702		3241		11921	

*Signifies superiority

Table 27. Summary of GEM-MOS Comparative Verification Results. Non-chi-square scores are expressed as fractions: numerator is the number of scores for which GEM is superior to MOS; denominator is the number of non-chi-square scores for an element. Chi-square scores are expressed as the number of stratifications (out of 4) for which the chi-square score for GEM is smaller than for MOS.

Element	Ceiling			Visibility			Total Cloud Amount		
Mode	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional
3-h Proj. Scores	4/4	1/4	4/4	4/4	2/4	4/4	3/3	0/3	3/3
# of Chi Sqs. Favoring GEM	0	0	0	1	0	1	1	2	1
9-h Proj. Scores	1/4	0/4	1/4	1/4	0/4	0/4	0/3	0/3	0/3
# of Chi Sqs. Favoring GEM	2	1	0	0	0	0	1	1	1

Element	Temperature			Dewpoint Depression			Wind		
Mode	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional
3-h Proj. Scores	4/5A	0/5	4/5B	2/2	0/2	2/2	2/2E	0/2	2/2F
# of Chi Sqs. Favoring GEM	1	0	1	3	0	2	0	0	0
6 or 9-h Proj. Scores	0/5	0/5	0/5	0/2C	0/2	0/2	0/2	0/2	0/2
# of Chi Sqs. Favoring GEM	0	0	0	2	1	0D	0	0	0D

Footnotes:

A - In stratifications, MOS favored over GEM in cool season/15 GMT GEM input time (12 GMT MOS cycle) stratification.

B - In one or more stratifications, MOS favored in the 12 GMT cycle (21 GMT GEM input time) stratifications.

C - GEM superior for warm season/03 GMT GEM input time (previous 12 GMT MOS cycle) stratification.

D - GEM chi-square values very much larger than MOS for 5 stratifications out of 24, for both dewpoint depression and wind.

E - MOS favored on warm season/15 GMT GEM input (previous 00 GMT MOS cycle) stratification.

F - MOS favored on 21 GMT GEM input time (12 GMT MOS cycle) stratifications for both seasons. GEM favored on 03 GMT GEM input time (00 GMT MOS cycle) stratifications for both seasons.

Table 28. Approximate crossover projection times (X) between GEM (G) and MOS (M) for the major aviation elements--ceiling, visibility, and total cloud amount. Results from operational, special operational, and scientific comparisons were used (see text). The circled letters indicate where information from the study's results were used.

Forecast Time		Projection (hours) 1	2	3	4	5	6	7	8	9	10	11	12
0000 GMT	1200 GMT	G	G	G	G	G	G	G	X	M	M	M	M
0100 GMT	1300 GMT	G	G	G	G	G	G	G	X	M	M	M	M
0200 GMT	1400 GMT	G	G	G	G	G	G	G	X	M	M	M	M
0300 GMT	1500 GMT	G	G	(G)	G	G	G	G	X	(M)	M	M	M
0400 GMT	1600 GMT	G	G	G	G	G	G	G	X	M	M	M	M
0500 GMT	1700 GMT	G	G	G	G	X	M	M	M	M	M	M	M
0600 GMT	1800 GMT	G	G	G	G	G	X	M	M	M	M	M	M
0700 GMT	1900 GMT	G	G	G	G	G	G	X	M	M	M	M	M
0800 GMT	2000 GMT	G	G	G	G	G	G	X	M	M	M	M	M
0900 GMT	2100 GMT	G	G	(G)	G	G	G	X	M	(M)	M	M	M
1000 GMT	2200 GMT	G	G	G	G	G	G	X	M	M	M	M	M
1100 GMT	2300 GMT	G	G	G	G	G	G	X	M	M	M	M	M

Table 29. Approximate crossover projection times between GEM and MOS for the lesser aviation elements--wind, temperature, and dewpoint depression. The circled letters indicate where information from the study's results were used.

Forecast Time		Projection (hours) 1	2	3	4	5	6	7	8	9	10	11	12
0000GMT	1200GMT	G	G	G	G	X	M	M	M	M	M	M	M
0100GMT	1300GMT	G	G	G	G	X	M	M	M	M	M	M	M
0200GMT	1400GMT	G	G	G	G	X	M	M	M	M	M	M	M
0300GMT	1500GMT	G	G	(G*)+	G	X	(M*)	M	M	(M)+	M	M	M
0400GMT	1600GMT	G	G	G	G	X	M	M	M	M	M	M	M
0500GMT	1700GMT	G	G	X	M	M	M	M	M	M	M	M	M
0600GMT	1800GMT	G	G	X	M	M	M	M	M	M	M	M	M
0700GMT	1900GMT	G	G	G	X	M	M	M	M	M	M	M	M
0800GMT	2000GMT	G	G	G	X	M	M	M	M	M	M	M	M
0900GMT	2100GMT	G	G	(G*)+	G	X	(M*)	M	M	(M)+	M	M	M
1000GMT	2200GMT	G	G	G	G	X	M	M	M	M	M	M	M
1100GMT	2300GMT	G	G	G	G	X	M	M	M	M	M	M	M

* Circled information for temperature and dewpoint depression
\+ Circled information for Wind

Table 30. Ceiling regression equations blending MOS and GEM. GEM uses 03 GMT and 15 GMT ceiling probabilities; MOS uses 15 GMT and 03 GMT probabilities, respectively, from the previous cycle.

ELEMENT: Ceiling
PROJECTION: 3 Hours

Predictors		Predictand Categories 1	2	3	4	5	6
Additive Constant		.001	-.001	.004	-.010	-.036	1.042
MOS Probability	1	.238	-.092	.027	.025	.250	-.441
for Predictor	2	-.040	.335	-.307	-.008	-.080	.100
Categories	3	.011	-.070	.649	-.024	.158	-.728
	4	.016	-.008	-.116	.310	.007	-.207
	5	-.013	.046	.012	-.077	.440	-.406
GEM Probability	1	.608	.267	-.082	-.131	-.029	-.634
for Predictor	2	-.031	.877	.294	-.107	-.098	-.934
Categories	3	-.008	-.060	.721	.078	-.102	-.630
	4	-.005	-.008	.005	.911	-.050	-.853
	5	-.000	-.012	-.027	-.047	.848	-.768

ELEMENT: Ceiling
PROJECTION: 9 Hours

Predictors		Predictand Categories 1	2	3	4	5	6
Additive Constant		-.002	-.002	-.004	-.013	-.039	1.061
MOS Probability	1	.611	.189	-.033	-.596	-.049	-.122
for Predictor	2	-.166	.237	-.104	.304	.141	-.412
Categories	3	.024	.233	.814	.081	.195	-1.346
	4	.016	-.062	.022	.604	.029	-.608
	5	-.001	.013	-.062	.042	.836	-.827
GEM Probability	1	.623	.371	-.250	-.135	-.491	-.119
for Predictor	2	.192	.821	.143	-.110	.028	-1.073
Categories	3	-.140	-.141	.602	-.392	-.206	.278
	4	-.007	-.014	-.044	.757	-.030	-.662
	5	.016	-.011	.013	-.158	.479	-.339

Table 31. Ceiling and visibility verification scores for MOS, GEM, and a blend of both. Percent improvement over MOS and GEM for the blend are displayed for all scores except chi-square, where the differences between MOS and blended, and GEM and blended, are displayed. GEM uses 03 GMT and 15 GMT probabilities; MOS uses 15 GMT and 03 GMT probabilities, respectively. The configuration of GEM and MOS corresponds to the special operational comparison of Section 3.

ELEMENT: Ceiling
PROJECTION: 3 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.166	.139	.135*	18.7	2.9
% Correct	72.9	79.5	79.8*	9.5	0.4
Heidke	.379	.474	.535*	41.2	12.9
Threat	.125	.325	.340*	172.0	4.6

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	5.86*	130.	6.17	-0.31	123.83

SAMPLE SIZE: 11244

ELEMENT: Ceiling
PROJECTION: 9 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.167	.177	.163*	2.4	7.9
% Correct	72.2	70.4	74.1*	2.6	5.3
Heidke	.362	.322	.393*	7.9	18.1
Threat	.134	.169	.185*	38.1	9.5

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	12.39	4.63*	6.38	6.01	-1.75

SAMPLE SIZE: 11243

Table 31 (continued)

ELEMENT: Visibility
PROJECTION: 3 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.113	.093	.091*	19.5	2.2
% Correct	81.4	86.2	86.4*	6.1	0.2
Heidke	.305	.492	.497*	63.0	1.0
Threat	.190	.351	.355*	86.8	1.1

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	8.12	10.0	3.07*	5.05	7.33

SAMPLE SIZE: 11284

ELEMENT: Visibility
PROJECTION: 9 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.139	.148	.135*	2.9	8.8
% Correct	76.7	72.7	78.5*	2.3	8.0
Heidke	.296	.224	.336*	13.5	50.0
Threat	.131	.157	.173*	32.1	10.2

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	6.88*	26.1	15.0	-8.12	11.1

SAMPLE SIZE: 11283

* Signifies superiority

Table 32. Ceiling and visibility verification scores for MOS, GEM and a blend of both. Percent improvement over MOS and GEM for the blend are displayed for all scores except chi-square, where the differences between MOS and blended, and GEM and blended, are displayed. GEM uses 09 GMT and 21 GMT probabilities; MOS uses 03 GMT and 15 GMT probabilities, respectively. The configuration of GEM and MOS corresponds to the operational comparison of Section 3.

ELEMENT: Ceiling
PROJECTION: 3 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.161	.135	.133*	17.4	1.5
% Correct	73.3	80.7T	80.7T	10.1	0.0
Heidke	.404	.509	.554*	37.1	8.8
Threat	.201	.345	.353*	75.6	2.3

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	5.75	100.0	4.15	1.60	95.85

SAMPLE SIZE: 11204

ELEMENT: Ceiling
PROJECTION: 9 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.170	.180	.166*	2.3	7.8
% Correct	72.0	70.0	74.1*	2.9	5.9
Heidke	.369	.331	.398*	7.8	20.2
Threat	.121	.158	.176*	45.5	11.4

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	7.65	7.20	17.5	-9.85	-10.30

SAMPLE SIZE: 11204

Table 32 (continued)

ELEMENT: Visibility
PROJECTION: 3 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	OVER GEM
Brier Score	.134	.119	.114*	14.9	4.2
% Correct	78.1	83.5*	83.4	6.8	-0.1
Heidke	.330	.467	.476*	44.2	1.9
Threat	.202	.319	.325*	60.9	1.9

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	(GEM-BLENDED)
Chi Square	15.3*	45.4	24.2	-8.9	21.2

SAMPLE SIZE: 11243

ELEMENT: Visibility
PROJECTION: 9 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	OVER GEM
Brier Score	.114	.122	.113*	0.9	7.3
% Correct	81.3	79.9	82.7*	1.7	3.5
Heidke	.301	.238	.327*	8.6	37.4
Threat	.175	.167	.250*	42.9	49.7

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	(GEM-BLENDED)
Chi Square	8.44*	35.6	19.2	-10.76	16.40

SAMPLE SIZE: 11243

* Signifies superiority
T Signifies Tie

Table 33. Temperature mean absolute errors resulting from application of different forms of feedback. See text for discussion of feedback types. Results are expressed for each of the 21 stations in the GEM-MOS comparative verification sample, as well as for the sample as a whole.

(1) Station	(2) Sample Size	(3) GEM Without Feedback	(4) GEM With Station Bias Feedback	(5) GEM With Type a Feedback*	(6) GEM With Type b Feedback*	(7) GEM With Type c Feedback*	(8) MOS Without Feedback	(9) MOS With Feedback
1 ABQ	112	3.96	3.96	3.32	3.40	3.25	2.62	2.60
2 BIS	81	5.56	5.43	4.74	4.73	4.58	3.37	3.35
3 BOS	108	4.98	4.93	4.91	4.87	4.87	2.93	2.92
4 BUF	89	3.66	3.61	3.67	3.63	3.60	2.94	2.95
5 DEN	105	4.05	3.78	3.71	3.52	3.52	3.29	2.94
6 LAX	111	7.13	2.85	4.64	2.59	2.57	1.94	1.80
7 MIA	44	2.86	2.86	2.85	2.81	2.84	2.30	2.29
8 MKE	99	4.10	4.11	4.22	4.15	4.07	3.70	3.55
9 MSP	109	4.72	4.72	4.42	4.20	4.20	3.50	3.51
10 MSY	79	3.47	3.18	3.35	3.16	3.13	3.14	2.86
11 OKC	97	5.19	4.87	4.17	4.13	3.84	3.01	2.86
12 PDX	94	4.54	4.42	4.04	4.07	4.06	2.88	2.83
13 PHL	106	3.92	3.66	3.44	3.40	3.40	3.25	2.64
14 RDU	106	3.08	2.99	2.77	2.77	2.77	2.52	2.28
15 SAT	89	3.65	3.78	3.35	3.45	3.45	3.37	3.10
16 SAV	103	3.73	3.76	3.57	3.57	3.57	3.07	2.94
17 SFO	93	4.44	3.85	3.84	3.60	3.58	3.10	3.11
18 SLC	84	4.96	4.56	4.29	4.13	4.10	2.52	2.46
19 STL	91	4.22	3.77	3.83	3.55	3.55	3.10	2.84
20 TYS	124	3.81	3.80	3.93	3.90	3.80	3.60	3.40
21 DCA	95	4.23	3.45	2.97	2.77	2.17	3.36	3.01
Total	2018							
Weighted Average		4.34	3.94	3.84	3.65	3.58	3.04	2.88

*See Figure 5 and text

Table 34. Weightings for GEM forecasted probabilities with Mahalanobis-distance *a posteriori* probabilities.

GEM forecasted probabilities	Mahalanobis-distance *a posteriori* probabilities
0.50	0.50
0.90	0.10
0.10	0.90

Table 35. Weightings for station-specific hourly and monthly climatologies.

Station-specific hourly climatology	Station-specific monthly climatology
0.50	0.50
0.67	0.33
0.33	0.67

Table 36. GEM wind comparative verification scores between OLD and NEW categorical selection procedures. Upper half of table displays data for GEM forecasts made from 09 GMT and 21 GMT observations; lower half of table displays data for GEM forecasts made from 03 GMT and 15 GMT observations.

OLD=Maximum probability categorical selection procedure.
NEW=Unaccumulated P-Star categorical selection procedure.

ELEMENT: Wind
PROJECTION: 3 Hours

Operational Comparisons

	GEM Input Obs Time 0900 GMT				GEM Input Obs Time 2100 GMT			
	Season				Season			
	Warm		Cool		Warm		Cool	
	OLD	NEW	OLD	NEW	OLD	NEW	OLD	NEW
% Correct	38.4	38.8*	39.7	39.8*	29.6*	29.2	27.6	28.0*
Chi Sq.	39	26*	33	27*	1115	103*	255	233*
Heidke	.322	.327*	.346	.347*	.249*	.245	.229	.233*
Sample Size	3031		3250		2380		3245	

Scientific Comparisions

	GEM Input Obs Time 0300 GMT				GEM Input Obs Time 1500 GMT			
	Season				Season			
	Warm		Cool		Warm		Cool	
	OLD	NEW	OLD	NEW	OLD	NEW	OLD	NEW
% Correct	34.5	34.9*	36.6	36.9*	29.6	29.8*	32.0	32.2*
Chi Sq.	91	75*	66	54*	35	25*	51	45*
Heidke	.285	.289*	.316	.318*	.241	.243*	.274	.276*
Sample Size	2788		3246		2691		3252	

*Signifies superiority

Table 37. GEM wind comparative verification scores between OLD and NEW categorical selection procedures. Upper half of table displays data for GEM forecasts made from 09 GMT and 21 GMT observations, lower half of table displays data for GEM forecasts made form 03 GMT and 15 GMT observations.

OLD = Maximum probability categorical selection procedure.
NEW = Unaccumulated P-star categorical selection procedure.

ELEMENT: Wind
PROJECTION: 9 Hours

Operational Comparisons

	GEM Input Obs Time 0900 GMT				GEM Input Obs Time 2100 GMT			
	Season				Season			
	Warm		Cool		Warm		Cool	
	OLD	NEW	OLD	NEW	OLD	NEW	OLD	NEW
% Correct	19.5*	18.1	17.9*	16.8	23.2*	22.9	18.8	20.0*
Chi Sq.	554*	561	792	673*	225	140*	335	128*
Heidke	.141*	.129	.124*	.119	.147*	.146	.117	.128*
Sample Size	3031		3250		2380		3226	

Scientific Comparisons

	GEM Input Obs Time 0300 GMT				GEM Input Obs Time 1500 GMT			
	Season				Season			
	Warm		Cool		Warm		Cool	
	OLD	NEW	OLD	NEW	OLD	NEW	OLD	NEW
% Correct	24.0	24.4*	23.6	24.2*	24.4*	23.0	18.2*	17.6
Chi Sq.	141	92*	244	149*	430	275*	600	369*
Heidke	.160	.166*	.167	.174*	.178*	.176	.123	.123
Sample Size	2788		3246		2691		3252	

*Signifies superiority

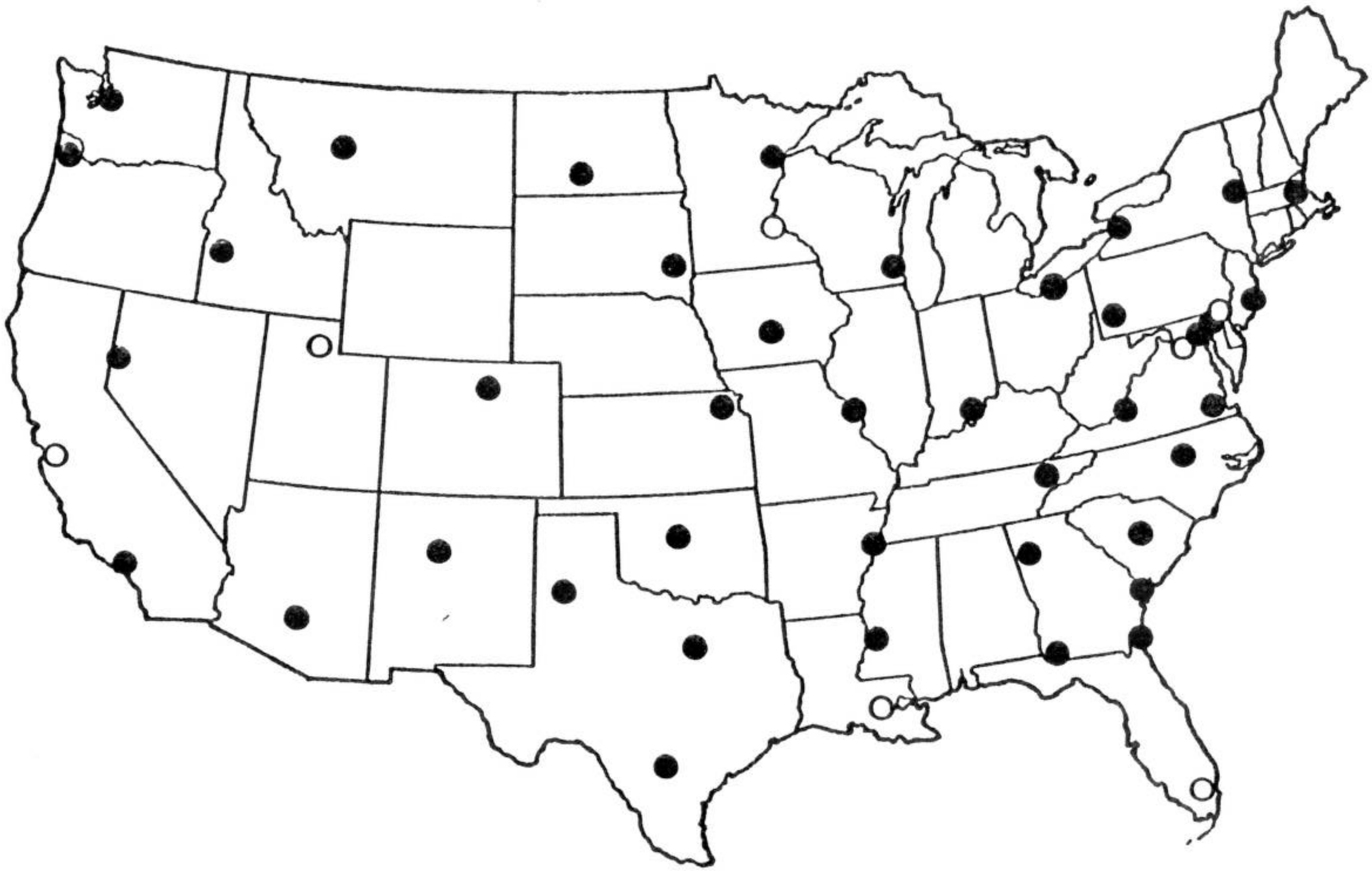

Figure 1. Stations used in developing GEM are shown as filled-in circles. Stations used to verify GEM against persistence are represented by open circles.

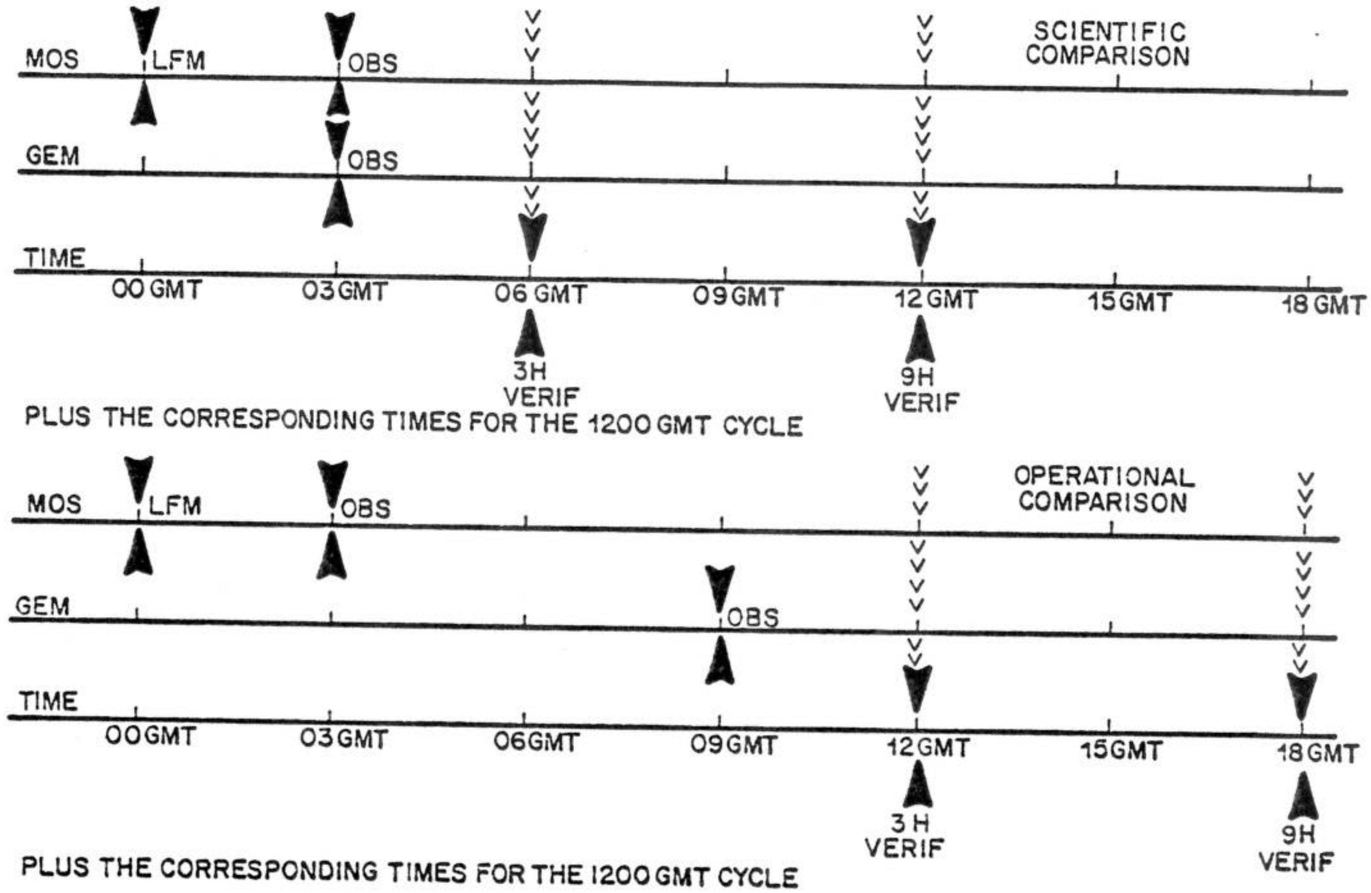

Figure 2. GEM-MOS comparative verification time lines. The upper part of the figure shows the scientific comparison; the lower part shows the operational comparison. Each part indicates the time of the model output (LFM), the surface observations (OBS) and the 3- and 9-h verification times used. The special operational comparison differs from the operational comparison as follows: The MOS and GEM input observation times differ by 12 hours instead of 6, and the MOS LFM predictors came from the previous model cycle, rather than the current cycle.

Figure 3. Stations used in the GEM-MOS comparative verification. Stations with 5-pointed stars were also used in a verification of GEM against persistence.

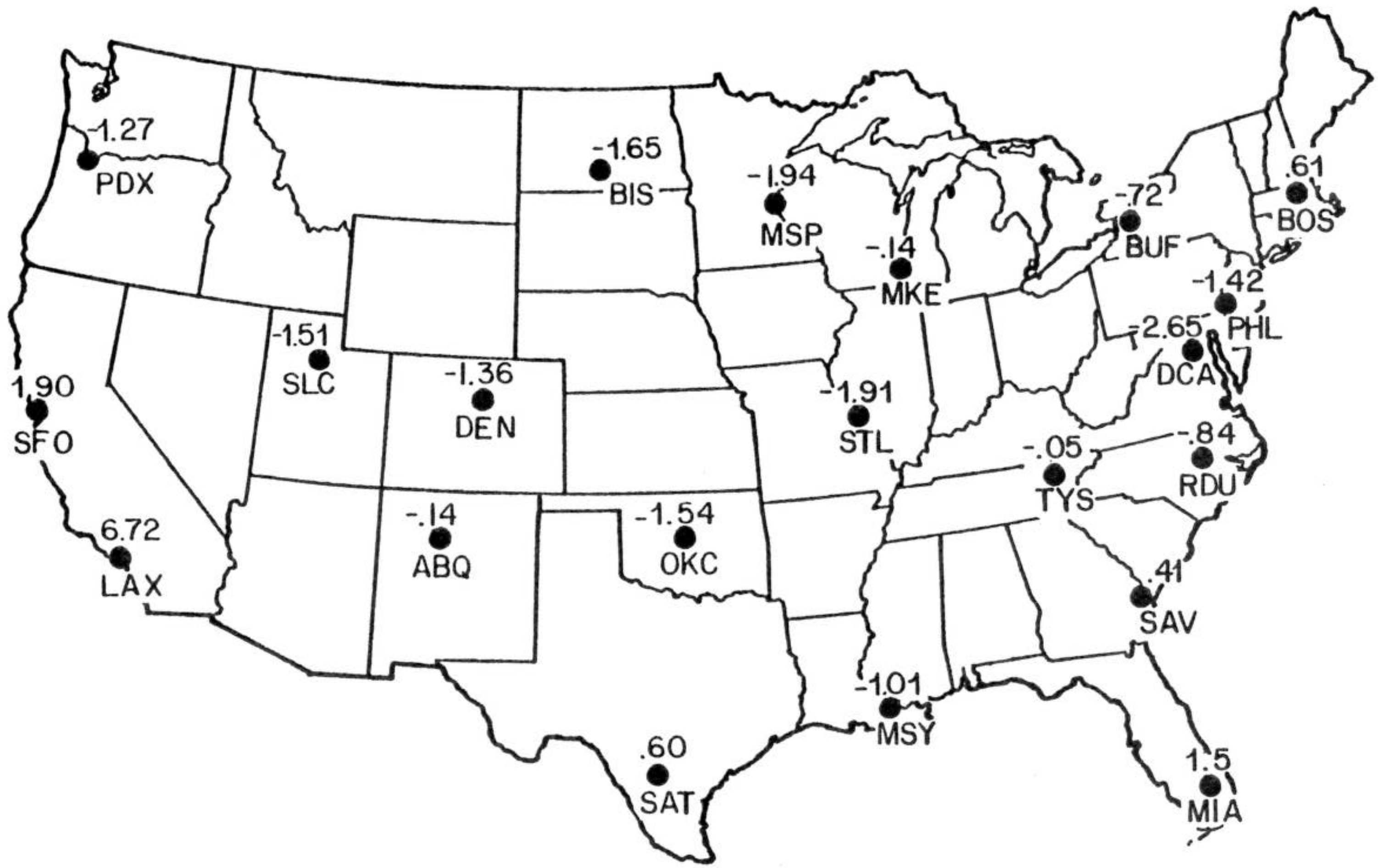

Figure 4. GEM mean algebraic error for 21 stations in the GEM-MOS comparative verification sample.

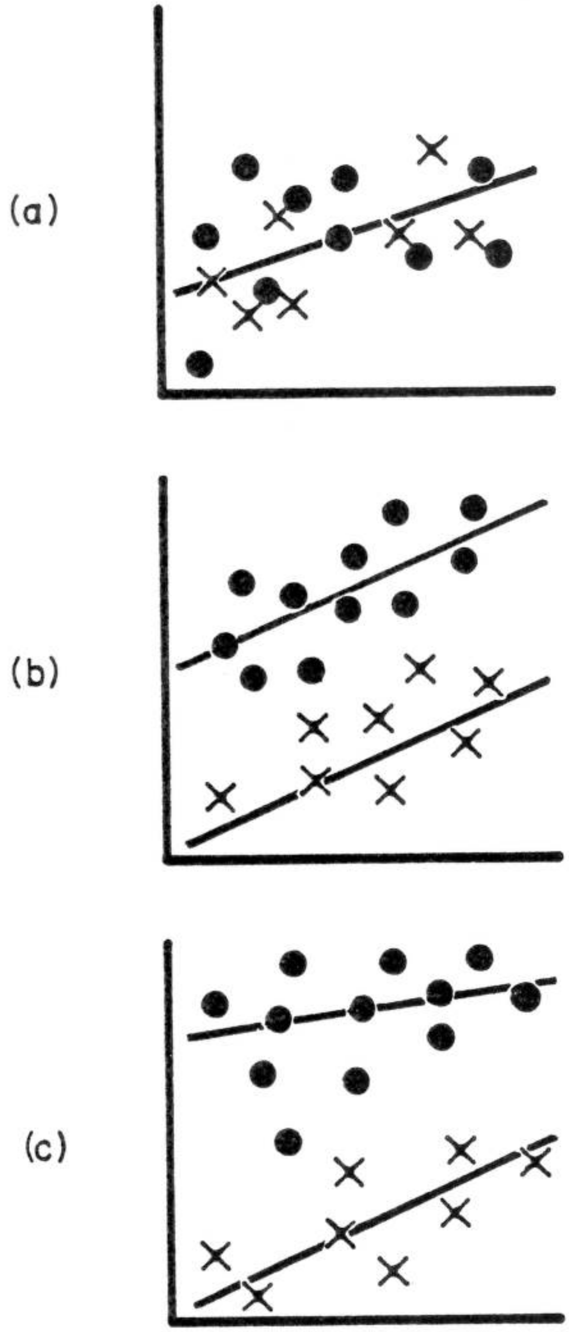

Figure 5. Feedback models used with temperature.

(a) Data from all stations are grouped together and a single regression line has been fitted to these data.

(b) A single slope is derived for all stations' data taken together, but a separate intercept is determined for each station.

(c) Both slope and intercept have been individually fitted to each station's data.