Approaches for Verification of Ceiling and Visibility Diagnoses and Forecasts

Barbara G. Brown[1], Jennifer L. Mahoney[2], Tressa L. Fowler[1], and Judy Henderson[2]

Quality Assessment Group
Aviation Forecast and Quality Assessment (AFQA) Product Development Team

31 August 2001

[1] Research Applications Program, National Center for Atmospheric Research (NCAR), Boulder, CO
[2] Forecast Systems Laboratory, National Oceanic and Atmospheric Administration (NOAA), Boulder, CO

1. Introduction

The purpose of this report is to describe verification approaches and methods that can be used for verification of national-scale ceiling and visibility forecasts and diagnoses. Some of these approaches are already being used in ongoing verification of operational forecasts, referred to as the Airmen's Meteorological Advisories (AIRMETs) of instrument-flight-rule (IFR) conditions (i.e., IFR AIRMETs) issued by the Aviation Weather Center (AWC), by the Forecast System Laboratory's Real-Time Verification System (RTVS; Mahoney et al. 1997). In addition, some of the methods are being used for ongoing feedback to the current automated product. The report considers general concepts of verification, as well as specific approaches that can be applied for various types of predictors and predictands (i.e., categorical, probabilistic, continuous). Results of a recent study regarding methods for matching forecasts and observations are also summarized, and the report on this topic is included in the Appendix.

2. General concepts

Some of the important concepts underlying the verification of forecasts are briefly considered here. These concepts are very important to keep in mind when designing verification systems and studies. Many of the ideas considered here are also presented in the verification literature, and are described well in such resources as Murphy (1997) and Wilks (1995). First and foremost, verification approaches should be statistically sound and scientifically meaningful.

2.1 Matching forecasts and observations

The initial step of creating a matched set of forecasts and observations can be one of the most difficult aspects of forecast verification. This process can be particularly difficult in the case of verification of aviation forecasts, where the observations frequently are non-standard. However, in order for verification results to be valid and meaningful, it is critical that the forecast and observed events are matched as closely as possible, in terms of time domain, spatial representativeness, and so on. In the case of ceiling and visibility, for example, a ceiling forecast might represent the ceiling at a particular grid point at a particular time. Since ceiling is observed only at particular airports and other locations, the matched set of forecasts and observations would be limited to those locations.

2.2 Verification framework

The statistical basis for forecast verification is the joint distribution of forecasts and observations, $p(f,x)$, where f represents the forecasts and x represents the observations (Murphy and Winkler 1987). This distribution can be decomposed into two conditional distributions (the conditional distribution of the forecasts given the observations and the conditional distribution of the observations given the forecasts) and two marginal distributions (the distribution of the forecasts and the distribution of the observations). These distributions form the basis for essentially all of the summary and performance measures that are generally used for verification (Murphy et al. 1989; Murphy and Winkler 1992). For example, bias is the difference between the

mean forecast and the mean observation, which are summary measures of the marginal distributions of forecasts and observations, respectively.

2.3 Dimensionality and the selection of appropriate measures

The joint distribution of forecasts and observations contains all of the non-time-dependent information about the quality of the forecasts (Murphy 1991). Thus, the "dimensionality" of the verification problem is the dimension of this distribution, which can be computed as $d = n(f)n(x) - 1$, where $n(f)$ is the number of possible forecasts and $n(x)$ is the number of possible observations. For example, for probabilistic forecasts of a Yes/No element (e.g., ceiling less than 10,000 ft), where 11 different probability values can be used, $n(f) = 11$, $n(x) = 2$, and $d = 21$. Thus, in this case, 21 different numbers are required to reconstruct $p(f,x)$. This result suggests that it is very desirable to consider a variety of measures when evaluating the quality of a set of forecasts, to respect the dimensionality of the forecast problem. Use of a single verification measure to evaluate a set of forecasts generally is not meaningful.

Of course, dimensionality is not the only consideration when selecting verification measures. Ideally, the choice of measures should be guided by the questions about forecast quality that are of interest. For example, if overall bias is of concern, then the mean error (ME) should be computed; if accuracy is of interest, then measures such as mean absolute error (MAE) and root mean squared error (RMSE) should be computed. In particular, different verification statistics measure different *attributes* of the quality of the forecasts. Use of this approach to verification represents a diagnostic verification of forecasts, which is much more informative than verification based on one or two "standard" measures or a single skill score. The actual statistics used to measure the various attributes depend on characteristics of the forecasts – i.e., whether they are continuous, categorical, or probabilistic.

Other factors also should be considered when selecting measures for a verification analysis, and particularly for long-term verification studies. For example, in some cases (e.g., for verification of dichotomous – e.g., Yes/No – forecasts) it is possible to select combinations of measures that can identify superiority of one forecasting system over another (see Section 2.6). It also is desirable to use measures that do not encourage a forecasting system to over- or under-forecast.

2.4 Standards of comparison

By nature, verification is a comparative process. In general, the specific verification values associated with a forecasting system are not meaningful or useful, without comparison to values associated with some other forecasting system or standard-of-comparison. Standards-of-comparison can be used to compute skill scores, which measure the relative improvement of one forecasting system over another. An appropriate standard of comparison could be another forecasting system (e.g., the current operational forecasts) or it could be based on a basic standard such as climatology or persistence. For ceiling and visibility forecasts, the operational forecasts (e.g., IFR AIRMETs; TAFs), persistence, and climatology all appear to be reasonable standards-of-comparison.

2.5 Relationships among measures

Because essentially all verification measures are derived from the joint distribution of forecasts and observations, it is not surprising that in many cases the different measures are strongly related. For example, the mean-squared error (MSE) can be broken down into several components, including the square of the mean error (ME) – thus, MSE can easily be dominated by the overall bias of the forecasts. As another example, the Critical Success Index (CSI; also known as the Threat Score), which is commonly used to summarize the skill of categorical forecasts, can be decomposed into a non-linear combination of the Probability of Detection (POD) and the False Alarm Ratio (FAR), as shown in Fig. 1. This figure illustrates that the impacts of changes in FAR and POD on the value of the CSI strongly depend on the values of POD and FAR (i.e., where the values lie on the curves). CSI and other skill scores and measures, including FAR, also are strongly influenced by the climatology, $p(x)$ – i.e., the frequency of occurrence of the event of interest (Brown and Young 2000; Mason 1989). Thus, these measures should not be used when the climatology is variable or not estimable, and they cannot be used to make comparisons between forecasting systems when the verification analyses are based on different time periods or locations with different climatologies.

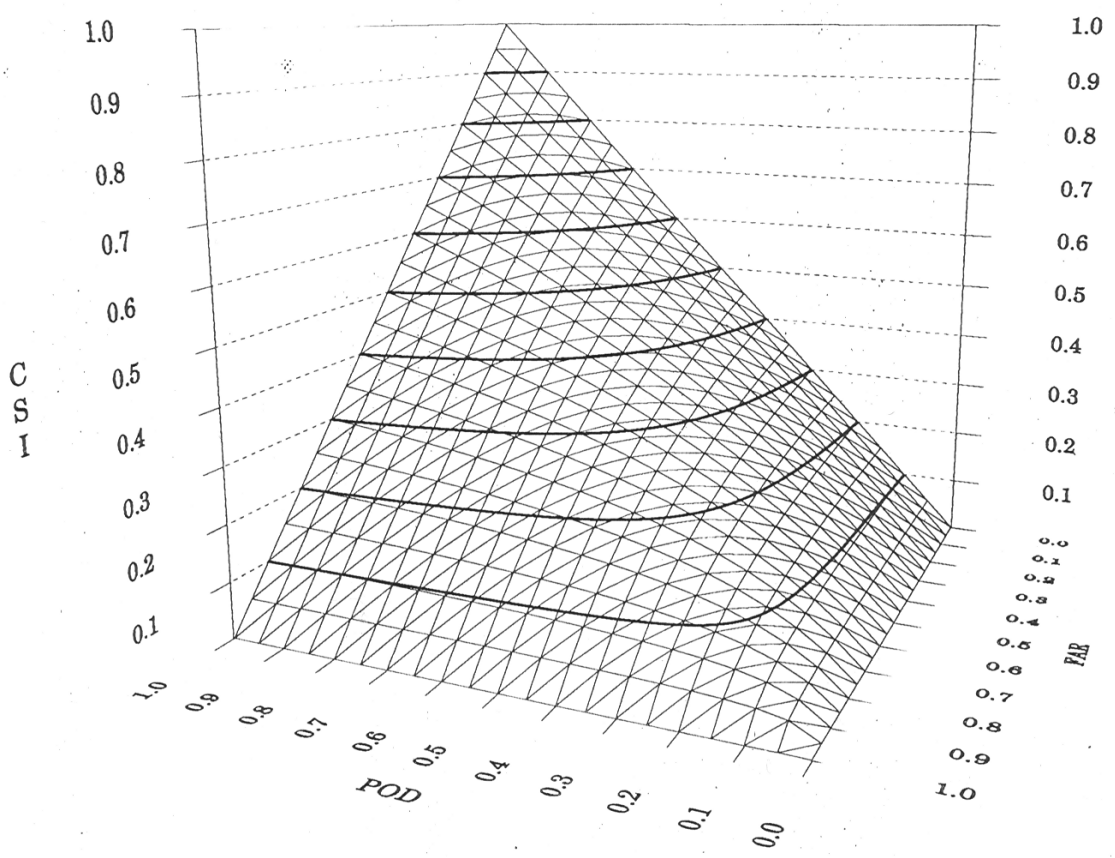

Figure 1. Critical Success Index (CSI) as a function of Probability of Detection (POD) and False Alarm Ratio (FAR).

2.6 Quality does not equal value

A common misconception that should be avoided is the idea that a forecasting system with a larger skill score also is more useful and possibly has greater economic value than a system with a lower skill score. Unfortunately, quality is not equivalent to value. In general, economic value is a non-linear function of quality; however, for some forecast users greater skill may not lead to greater value (for example, if some other forecast attribute is of greater importance than the measure that is being used to compare the forecasting systems). In general, assessment of the value of a forecasting system requires detailed examination and modeling of specific decision-making situations.

In some limited cases (as noted in Section 2.3), it is possible to state that one forecasting system is superior for all users. This conclusion can be reached in the case of verification of dichotomous forecasts, where certain combinations of measures are sufficient to make this comparison between forecasts. Examples of these combinations of measures are POD and FAR, POD and PODn (the probability of detection associated with "No" events), and POD and *p(x)*.

In recent years, some methods have been developed to translate verification measures for probabilistic forecasts into measures of the economic value of the forecasts (e.g., Richardson 2000; Wilks 2001). These approaches are based on a relatively simplistic decision-making model known as the cost-loss model. Essentially, the approaches allow computation of the relative value that would be attained by users of the forecasts with different cost-loss ratios. Although the approach is relatively simplistic, it provides a basic measure of potential economic value that should be of interest to researchers and program managers. The availability of these approaches provides another incentive for formulation of probabilistic forecasts.

2.7 Demonstrating true improvements in forecasting systems

Demonstrating superiority of one forecasting system is a desirable result of verification, when evaluating improvements to a forecasting system. However, it is important to recognize that this goal generally is *very* difficult to achieve. A common frustration is the fact that improvements in one measure of quality generally are associated with degradations in another measure (e.g., the common trade-off between POD and FAR).

A related issue that will be of concern in verification of national-scale ceiling and visibility products is the stringency that is associated with verification of forecasts on a grid, at least as they are commonly verified. In particular, standard approaches – in which forecasts at particular points are compared to observations at those points – do not take into account minor errors in location or timing. Forecasts are penalized equally for small errors and large errors. It will be desirable to investigate and develop alternative approaches that are more diagnostic and informative.

2.8 Other issues

The quality of forecasts generally varies from day-to-day. Moreover, as noted earlier, some verification statistics vary according to the frequency of the forecasted "event," which also can vary a great deal from day-to-day (e.g., the occurrence of low ceilings is more common on days that are strongly influenced by frontal systems). Thus, it is important to examine variations of the verification statistics from day-to-day, as well as considering long-term values of the measures across a season or other period. Experience with other types of forecasts has indicated that the verification statistics for some forecasting systems can be less variable than the statistics for other forecasting systems. These differences can be important to users; for example, having a consistent forecasting system may be more important to some users than having a better overall score. That is, it may be more important than having a system that occasionally is right on the mark, which also sometimes completely misses the mark.

Variability also is an important factor in defining confidence intervals for the verification statistics. Confidence intervals can help to determine if the verification statistics for two different forecasting systems are significantly different from each other (e.g., whether the values for one system are significantly better than the values for another system). Basic confidence intervals can be defined for statistics based on verification of dichotomous forecasts (Seaman et al. 1996) and these approaches have been extended to forecasts with more complicated sampling problems (Kane and Brown 2000). Extensions to other types of forecasts (e.g., continuous, probabilistic) also should be investigated.

One issue that is of particular concern for verification of ceiling and visibility forecasts is the fact that these phenomena generally are not continuous in spatial extent. This issue is related to the problems associated with the stringency of common verification approaches. Some investigation and testing of methods will be required to develop approaches that appropriately take this variability into account.

3. Matching forecasts and observations

At least initially, METARs will provide the basic data for verification of the national-scale ceiling and visibility products. METARs generally provide observations of ceiling and visibility conditions once per hour. However, special reports are sometimes issued (e.g., when conditions are changing), and not all stations report consistently. Thus, the number of reports varies from hour-to-hour as stations report or not. Depending on the consistency of reports, it may be desirable to compute statistics based on (a) all available stations and (b) a subset of stations that reports consistently (to provide a stable dataset).

To start with, the matching of forecasts and observations will be "driven" by the METAR stations in order to avoid matching grid point forecasts to METARs that are located far from the gridpoint. In particular, the observations at each METAR location will be matched to the forecasts from the nearest grid point. It also will also be of interest to consider additional gridpoints (e.g., the four closest gridpoints), to consider the impacts of local variability on the verification results.

The approach described above is based on evaluations at individual METAR locations. However, the METAR stations are not distributed evenly across the continental United States. In fact, some regions have a dense network of stations, whereas other regions are sparsely covered. The matching approach described above thus places greater weight on regions with dense coverage. An alternative approach was investigated in which a grid is overlaid on the observations. With this approach, the stations and forecasts within a grid area are used to assign a single forecast and observation to that grid area. Verification statistics associated with each of the methods were compared in a study of IFR AIRMET verification statistics, which is described in the Appendix. The results indicated that the verification statistics are relatively insensitive to the choice between these two approaches: both POD and FAR were only decreased a small amount when the gridded method was used.

4. Verification measures and approaches

As was indicated earlier, appropriate verification approaches and measures depend on the definition of the forecast event. In particular, forecast attributes are defined differently depending on whether the forecasts are continuous, categorical, or probabilistic. We expect, however, that the national-scale ceiling and visibility forecasts will be formulated in all three ways, at some point in the development process. Currently the system produces continuous values of ceiling and visibility, and it is important to evaluate these basic forecast values. Users are interested (for example) in the question of IFR vs. non-IFR conditions, and they may also be interested in specific categories of ceiling and visibility. Moreover, the operational forecasts (e.g., IFR AIRMETs) essentially are categorical. Thus, it is important to evaluate the forecasts as categorical predictions. Finally, we anticipate that it will be desirable in the future to provide probabilistic forecasts of ceiling and visibility categories. Thus, it is important to consider appropriate verification measures for probabilistic forecasts.

4.1 Methods for continuous variables

Verification of continuous predictands with corresponding continuous observations generally involves a fairly direct comparison of forecasts and observations. Attributes that are of interest include overall bias (measured by ME) and accuracy (measured, for example, by MAE and RMSE). Recall that RMSE is partially a function of ME. Other components of this measure include the variance and covariance of the forecasts and observations. The correlation coefficient frequently is computed for these types of variables; however, this measure ignores biases in the forecasts. Skill scores can be computed based on the RMSE (or other measures) using the standard formula for a skill score. The skill score essentially measures the percentage improvement of one forecasting system over another, relative to the value associated with perfect forecasts. Approaches that are more diagnostic include examination and display of the conditional quantiles of observations given forecasts (e.g., Murphy et al. 1992). These statistics represent changes in the distributions of observations as the forecast changes. Of course, the scatterplot is the simplest – and one of the most informative – displays representing the quality of continuous forecasts.

4.2 Methods for categorical forecasts

Categorical forecasts indicate that a particular category of observation will occur, such as Rain/No Rain. Categorical forecasts can either be based on ordinal categories (e.g., for cloud amount) or nominal categories (e.g., Yes/No). In the simplest (and most common) case there are two categories of forecasts and two categories of observations (i.e., the 2x2 case). Most measures that have been developed to evaluate categorical forecasts are designed for the 2x2 case; however most of them can be generalized to the n x m case either directly or by subdividing the table and combining categories to create various sets of 2x2 tables. Verification of categorical forecasts is based on examination of the frequencies of occurrence of various pairs of forecasts and observations (e.g., Table 1). These counts represent the joint distribution of forecasts and observations.

Table 1. Basic contingency table for evaluation of dichotomous (e.g., Yes/No) forecasts. Elements in the cells are the counts of forecast-observation pairs.

	Observation		
Forecast	***Yes***	***No***	***Total***
Yes	YY	YN	YY+YN
No	NY	NN	NY+NN
Total	YY+NY	YN+NN	YY+YN+NY+NN

Some of the verification measures that are available for evaluation of categorical forecasts (and which may be appropriate for evaluation of ceiling and visibility forecasts) are listed in Table 2. PODy and PODn are estimates of the proportions of *Yes* and *No* observations that were correctly forecasted, respectively. Together, PODy and PODn measure the ability of the forecasts to discriminate between *Yes* and *No* observations. The True Skill Statistic (TSS) (Doswell et al. 1990), also known as Hanssen-Kuipers discrimination statistic (Wilks 1995), summarizes this discrimination ability. Note, however, that it is possible to obtain the same value of TSS for a variety of combinations of PODy and PODn. Thus, it always is important to consider both PODy and PODn along with TSS**.** FAR estimates the frequency of Yes forecasts did not verify. CSI is the proportion of correct Yes forecasts, relative to the number of times the event was forecasted to occur or occurred. As noted earlier (Section 2.5), CSI is a nonlinear function of PODy and FAR. Like the TSS it should not be considered alone, without also examining PODy and FAR. One unfortunate aspect of the CSI is that it rewards over-forecasting. The Gilbert Skill Score (GSS) attempts to compensate for this effect by subtracting the number of correct Yes forecasts that would be expected to occur by chance. Similarly, the Heidke skill score (HSS) corrects the % Correct by subtracting the number that would be expected to be correct by chance. Finally, the % Area statistic is a measure of the forecast coverage. This statistic can be used as a surrogate indicator of over-warning.

Table 2. Verification measures for categorical forecasts.

Statistic	*Definition*	*Description*
PODy	YY/(YY+NY)	Probability of Detection of "*Yes*" observations
PODn	NN/(YN+NN)	Probability of Detection of "No" observations
TSS = HKSS	PODy + PODn – 1	True Skill Statistic; Hanssen-Kuipers Skill Statistic
FAR	YN/(YY+YN)	False Alarm Ratio
CSI = TS	YY/(YY+NY+YN)	Critical Success Index; Threat Score
GSS=ETS	(YY – C1)/(YY+NY+YN-C1)	Gilbert Skill Score; Equitable Threat Score
HSS	(YY+NN-C2)/(YY+YN+NY+NN-C2)	Heidke Skill Score
% Area	(Forecast Area) / (Total Area) x 100	% of the area of the continental U.S. where IFR conditions are forecast to occur

4.3 Methods for probabilistic forecasts

Categorical forecasts can be considered to be probabilistic forecasts that are completely certain. In this case, the forecasts can take on probability values between 0 and 1. In some situations, such as for human forecasters, the forecast values are limited to a discrete set of specific values. In contrast, the observations take on only the values 0 and 1 [i.e., either the event occurs ($x = 1$) or it does not occur ($x = 0$)]. Probabilistic forecasts can also be associated with multiple categories (e.g., varying categories of cloud cover).

Accuracy of probabilistic forecasts can be measured using the Brier score, which is a squared error measure like the MSE. Thus, this score can be decomposed in much the same way as the MSE. The Brier Score typically is translated into a skill score, generally with the sample climatology as a standard of reference. A comparable score – the Ranked Probability Score – is also available for multi-category probabilistic forecasts. This score can also be decomposed into more basic elements.

Another approach to verification of probability forecasts is based on Signal Detection Theory (SDT). The underlying basis for this approach is to select different probability thresholds to define Yes/No forecasts and to evaluate the resulting categorical forecasts associated with each threshold. The curve joining the (1-PODn, PODy) points for different probability values is known as the "Relative Operating Characteristic" (ROC) curve in SDT. The goal is for the ROC curve to lie close to the upper left corner of the diagram. The area under this curve is a measure of overall forecast skill (e.g., Mason 1982), and provides another measure that can be compared

among various forecasts. A forecast with no skill has an ROC area of 0.5 or less. One concern associated with use of the ROC is that it does not penalize forecasts that are not calibrated or reliable.

Forecast calibration can most easily be examined through a reliability diagram, relating particular forecast probabilities (x-axis) to the frequency of occurrence of the event of interest on occasions when that probability was used (y-axis). The curve for completely reliable forecasts lies directly on the diagonal. A more complete picture of forecast quality is provided by the "Attributes" diagram, which includes reliability as well as a variety of other measures related to the joint, conditional, and marginal distributions (Wilks 1995). An opposing diagram – the discrimination diagram – can be used to examine the distribution of forecast probabilities given particular observations (i.e., Yes and No; e.g., Murphy and Winkler 1992).

5. Current RTVS verification approaches for IFR AIRMETs

The verification approaches that have been used in RTVS to evaluate the IFR AIRMETs since 1997 in general are relatively simple and straightforward. These basic approaches are described below. Enhancements to the approaches are desired. In particular, it will be desirable to develop a methodology for evaluating these forecasts so that the results are comparable to the results associated with the automated products.

5.1 Observations and forecasts

The METAR surface observations are used to indicate IFR and non-IFR conditions. No attempt is made to establish a consistent table of METAR reports. All stations that report each hour are used in the verification process.

The IFR AIRMET forecasts are areal text forecasts of IFR conditions, where ceilings less than 3,000 ft and/or visibility less than 3 miles are expected. These forecasts are not issued unless the areal extent of the forecast region exceeds 3,000 square miles. The polygon, defined by the from-line that is specified in the IFR AIRMET, is verified without the inclusion of specific details that are outlined in the free-form text portion of the AIRMETs. In the evaluation of the IFR AIRMETs by RTVS, the forecasts and observations are treated as dichotomous Yes/No forecasts.

5.2 Matching forecasts and observations

Within RTVS, the IFR AIRMETs are evaluated two ways: as an hourly forecast (observation-based verification) and as a 6-h forecast (forecast-based verification).

When IFR AIRMETs are evaluated as an hourly forecast, the METAR reports are examined hourly to determine whether the report meets AIRMET criteria for IFR conditions (ceilings less than 3,000 ft and/or visibility less than 3 miles). The METAR report is then evaluated to determine whether it falls within the AIRMET boundary. A single observation from one METAR station is used each hour to evaluate the AIRMET. This method is different from

the methods used to evaluate the 6-hourly AIRMETs, described below. If the METAR report falls within the AIRMET boundary and meets the AIRMET criteria, then a YY pair is recorded. If the METAR report falls outside the AIRMET boundary and meets AIRMET criteria, then a NY is recorded, and so on filling in the pairs of the 2x2 contingency table (Table 1).

When the IFR AIRMETs are verified as a 6-h forecast, only one METAR observation over the 6-h period is used to verify the AIRMET over the valid period. For instance, if METAR station *SLC* reports IFR conditions within 3 different hours over the 6-h AIRMET valid period, *SLC* is used only once as a Yes observation to verify the AIRMET. If *SLC* never meets IFR criteria within the 6-h AIRMET valid period, then only one No observation is recorded. This approach was designed to meet the needs of the AIRMET forecast.

A third approach for verifying the AIRMETs, where the METAR reports were gridded over the national domain for consistency, was tested in RTVS. The results from this test are described in the Appendix.

5.3 Statistics generated by RTVS

The statistics computed by RTVS are consistent with those summarized in Table 2. Displays and contingency tables are available through a Web-based interface that can be accessed from http://www-ad.fsl.noaa.gov/afra/rtvs; link to Real Time Verification System.

5.4 Issues

Several issues associated with the verification approaches used in RTVS are apparent. First, the details in the IFR AIRMET forecast are difficult to evaluate since the information is difficult, if not impossible, to decode. Second, since the forecasts are not consistently distributed throughout the forecast domain, the FAR values are lower than they would be if the forecasts were evenly distributed. Third, the observations are not smoothed to accommodate the 3,000 sq. ft AIRMET criteria. Ideally, development of methods to evaluate the automated products will also lead to improvements in the methods used for the IFR AIRMETs.

6. Issues

One aspect of the national-scale products that has not yet been addressed is the diagnosis, which currently is based to a large extent on the METAR observations. A key question relative to the diagnoses is how well they capture the correct ceiling and visibility in locations without observations. One approach to this issue is to create a large set of forecasts with one or more stations removed from the analysis. Then, the data for the stations that were removed can be used to verify the diagnoses. This issue also relates to a general concern that the observations have relatively little spatial continuity, so that the characteristics of the ceiling and visibility between stations are unknown. This uncertainty limits the scope of the verification that be accomplished without additional data.

It was noted earlier that it would be desirable to develop verification methods that are more diagnostic. Such methods would make the verification results more meaningful for forecast developers as well as managers. In addition, it would be desirable to begin investigating the use of remote-sensing (e.g., satellite) data and pilot reports as verification tools for ceiling and visibility. However, use of these types of data will require time to test the methods based on them.

7. Summary

This report has presented a smorgasbord of ideas for possible verification approaches that can be taken as the national-scale Ceiling and Visibility product matures. However, it is important to note that – just as algorithms and forecasting systems evolve and improve with time, so too do verification approaches and methods. Thus, development of verification methods and approaches to keep up with the algorithm development will require time to develop and test new methods. In addition, we view this as a "living document" that will evolve as our knowledge of the ceiling and visibility verification problem expands.

Acknowledgments

This research is in response to requirements and funding by the Federal Aviation Administration. The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government.

References

Brown, B.G., and G.S. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, American Meteorological Society (Boston), 393-398.

Doswell, C.A., III, R. Davies-Jones, and D.L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576-585.

Kane, T.L., and B.G. Brown, 2000: Confidence intervals for some verification measures – a survey of several methods. *Preprints, 15th Conference on Probability and Statistics in the Atmospheric Sciences*, Asheville, NC, 8-11 May, American Meteorological Society (Boston), 46-49.

Mahoney, J.L., J.K. Henderson, and P.A. Miller, 1997: A Description of the Forecast Systems Laboratory's Real-Time Verification System (RTVS). *Preprints, 7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, American Meteorological Society (Boston), J26-J31

Mahoney, J.L., J. Henderson, and B. Brown, 2001: A Comparison between Two Verification Approaches for Evaluating the IFR AIRMETs. Report, available from 1-5 on Thursdays.

Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291-303.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Australian Meteorological Journal*, **37**, 75-81.

Murphy, Allan H., 1991: Forecast verification: Its Complexity and Dimensionality. *Monthly Weather Review*, **119**, 1590–1601.

Murphy, A.H., 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts*. R.W. Katz and A.H. Murphy, Editors, Cambridge University Press, 19-74.

Murphy, A.H. and R.L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Murphy, A.H., B.G. Brown and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485-501.

Murphy, A.H., and R.L. Winkler, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435-455.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126**, 649-667.

Seaman, R., I. Mason, and F. Woodcock, 1996: Confidence intervals for some performance measures of Yes-No forecasts. *Australian Meteorological Magazine*, **45**, 49-53.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Wilks, D.S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209-219.

Appendix: Comparison of verification results for IFR AIRMETs, based on two matching approaches

APPENDIX

A Comparison between Two Verification Approaches for Evaluating the IFR AIRMETs

Jennifer Luppens Mahoney,

Judy K. Henderson,

and

Barbara Brown

28 June 2001

1. Introduction

Two different verification approaches for evaluating forecasts of ceiling and visibility Airmens' Meteorological Advisories (AIRMETs) were evaluated from 1 – 31 January 2001. The purpose for the evaluation was to determine whether the current station-by-station approach used in the Real-Time Verification System (RTVS) to evaluate the IFR AIRMETs was impacting the False Alarm Ratio (FAR) calculation because of the inconsistent spacing of the METAR stations. Therefore, in hopes to alleviate the inconsistent spacing problem, a gridded approach was applied to the RTVS and tested. Statistical verification results from the two approaches are summarized in this report.

2. Forecasts and Observations

The IFR AIRMETs, evaluated using the two verification approaches, are human generated text forecasts that are issued when ceilings are less than 1,000 ft and/or visibility is less than 3 miles. These forecasts are issued every 6-h, but are amended when weather conditions meet or exceed the AIRMET criteria. In this study, only the 6-h AIRMET forecasts were considered, because: 1) the methods, when applied to the 6-h AIRMETs, were less complicated than for the amended AIRMETs and 2) the relative differences in results should be similar for the amended forecasts.

Surface observations of ceiling and visibility (METAR reports) are used to verify the IFR AIRMETs. In particular, observations within the 6-h valid period of the AIRMET were used to verify the forecast. All observations, including specials, during the 6-h period were considered in the verification procedures.

3. Matching Approaches

In the current station-by-station approach, each METAR station is considered over the 6-h valid period of the AIRMET. If the observation meets the AIRMET criteria just once within the 6-h period and falls within an AIRMET boundary, then a YY (forecast/observation) pair is assigned to that observation. If the observation never meets the AIRMET criteria within the 6-h period and falls within the AIRMET boundary, then one YN pair is assigned, and so on, completing the standard 2x2 contingency table.

The second verification method involves mapping the METAR observations to a 40-km grid prior to generating the forecast/observation pairs. Each METAR station is assigned to a particular grid box within the national domain. In some instances, a grid box could contain as many as six METAR stations. Instead of using each individual METAR station to obtain the forecast/observation pair, a pair is generated for each grid box. For instance, if one observation within the grid box meets the AIRMET criteria and the grid box is included inside an AIRMET, a YY pair is assigned to that grid box. As a further example, if six METAR

stations fall within one grid box and 3 of the stations meet the AIRMET criteria, a single Yes observation is assigned to the grid box. This approach differs from the current station-by-station method in which 3 Yes observations and 3 No observations would be assigned to the grid box. A Yes forecast would be assigned to the grid box if any part of it is covered by an AIRMET. In this way, one of the possible forecast/observation pairs (YY, YN, NY, or NN) is assigned to each grid box.

4. Verification Measures

A summary of the verification measures used to evaluate the IFR AIRMETs is shown in Table 1.

Table 1. Verification statistics used in this study

Statistic	***Definition***	***Description***
PODy	YY/(YY+NY)	Probability of Detection of "Yes" observations
PODn	NN/(YN+NN)	Probability of Detection of "No" observations
FAR	YN/(YY+YN)	False Alarm Ratio
CSI	YY/(YY+NY+YN)	Critical Success Index
Bias	(YY+YN)/(YY+NY)	Forecast Bias
TSS	PODy + PODn – 1	True Skill Statistic
Heidke	[(YY+NN)-C1]/(N-C1), where N=YY+NY+NY+NN C1=[(YY+YN)(YY+NY) + (NY+NN)(YN+NN)] / N	Heidke Skill Score
Gilbert	(YY-C2)/[(YY-C2)+YN+NY], where C2=(YY+YN)(YY+NY)/N	Gilbert Skill Score

5. Results

The PODy and FAR values computed for the IFR AIRMETs (with no amendments) using the station-by-station verification approach (current method) are compared to the results generated using the gridded approach and are shown in Figs. 1 and 2, respectively. The daily statistics shown in the figures were computed from the forecast/observation pairs accumulated for all AIRMET issue times from 1 –31 January 2001. As shown by the results in the Figs., the differences in the PODy and FAR values associated with using the differing methods is small and confidence intervals for the values indicate the differences are not statistically significant, although some daily differences are larger than others.

Overall results computed for the 1-month period are shown in Table 2. Differences in the overall PODy, PODn, and FAR are very small, and the skill score (CSI, HSS, and TSS) values based on the two methods are almost identical. The PODy values actually decrease slightly, by less than 0.03, when the gridded method is applied. Similarly, the FAR values decrease for the gridded method, but by only 0.02. Differences between the CSI, TSS, and HSS values are 0.01 or less.

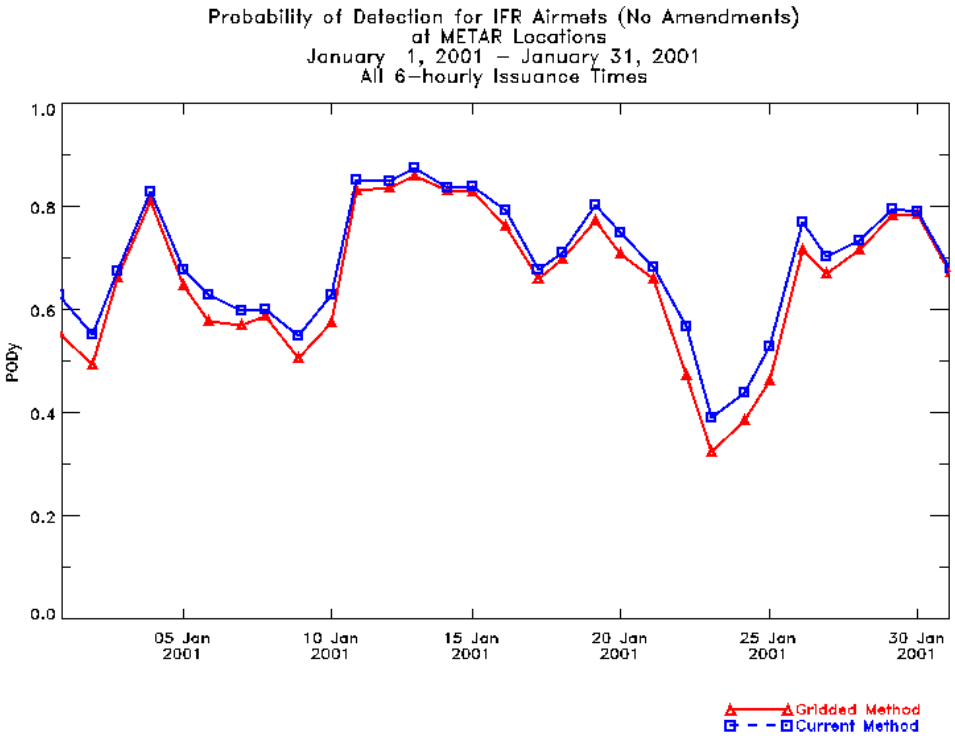

Fig. 1. Daily time series of PODy from 1-31 January 2001. Values computed using a station-by-station method ('squares') and a gridded approach ('triangles') are shown. Values were computed by accumulating pairs for all AIRMET issue times every day.

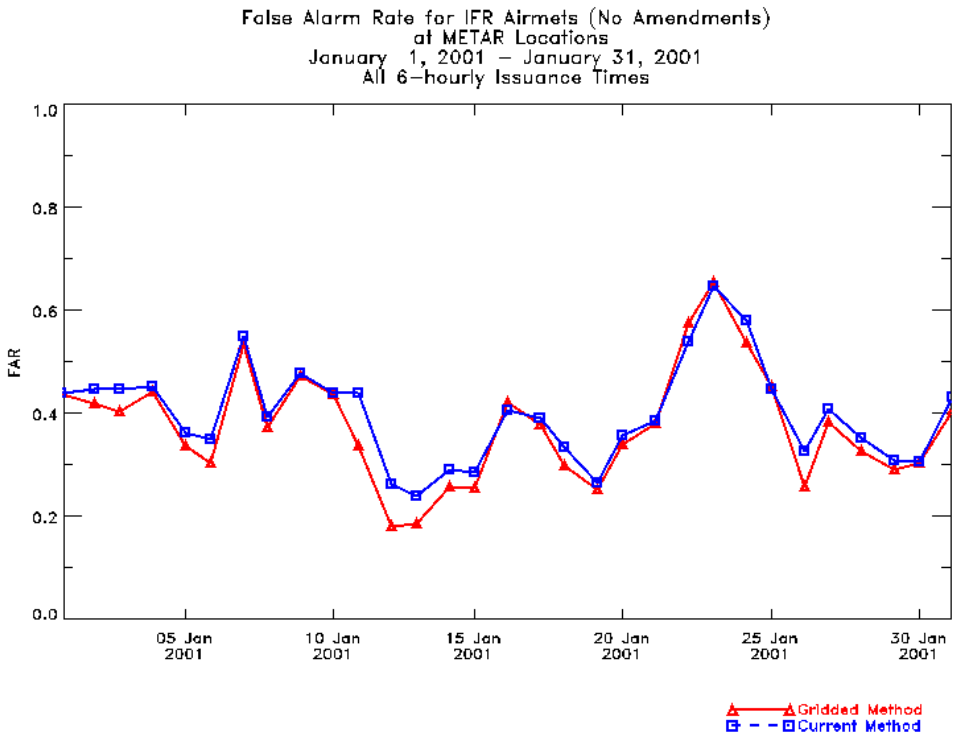

Fig. 2. Same as Fig. 1, except for FAR.

Table 2. Overall results for the IFR AIRMETs from 1-31 January 2001 generated using the current station-by-station verification method and the gridded method.

Method	**Statistic**						
	PODy	**PODn**	**FAR**	**CSI**	**TSS**	**HSS**	**Bias**
Current	0.73	0.86	0.37	0.51	0.59	0.56	1.2
Gridded	0.70	0.88	0.34	0.51	0.58	0.57	1.1

6. Recommendations

In summary, the differences in the statistical results for the IFR AIRMETs computed using the current station-by-station method and the gridded method are very small. Therefore, since the benefits of implementing the gridded approach do not out weigh the time and effort it takes to implement a new approach into RTVS (e.g., implementing the new approach system wide, setting up new real-time processes, making a new direction structure, gathering historical data, re-running all historical data to reflect the new approach, and changing the database and graphical user interface to reflect these changes), we recommend that the

current verification approach remain as the standard of comparison until another approach is proven to provide added benefit to the verification of the IFR forecasts.