APPENDIX A – Verification Scores

Table of Contents:

1. Introduction. Verification scores are applied at the local, regional, and national levels. Different scores may be applied to the same data. The type of score selected for use depends upon the objective. Frequently used scores are given in this manual and presented within the context of specific elements and events subject to verification. An excellent reference for verification scores is Wilks (1995).

In general terms, the scores are measures of accuracy and skill. **Accuracy** is a measure of how much a forecast agrees with the event or element being forecast. The smaller the difference between the forecast and observation, the greater the accuracy. **Skill** is a measure of improvement of a forecast over an established standard. Examples of standards often used for comparison include the climatological frequency (or value), persistence, or forecasts made by another process (e.g., model output statistics). The greater the improvement, the greater the skill.

2. Generalized Contingency Table. A generalized forecast/observation contingency table (Table A-1) is often used to summarize the forecast performance of a given element by category

(the term "category" is sometimes called class). The table is divided into *k* mutually exclusive and exhaustive categories. Each cell of the table, A_{ij}, gives the number of occurrences with the observation in the *ith* category (e.g., 13 to 17 knots for sustained wind speed) and the forecast in the *jth* category (e.g., 18 to 22 knots for sustained wind speed). Categorically correct forecasts (A_{ii} for all *i)*, where all $i = j$, fall along the upper left to lower right diagonal of the contingency table. The row and column totals, R_i and C_i, respectively, are often called the marginal totals of the contingency table, and they are used in computing forecast bias and skill.

Table A-1. Generalized Contingency Table

	Forecast Category				
Observed Category	1	2	…	k	Total
1	A_{11}	A_{12}	...	A_{1k}	R_1
2	A_{21}	A_{22}	...	A_{2k}	R_2
…	...	...	...	...	...
k	A_{k1}	A_{k2}	...	A_{kk}	R_k
Total	C_1	C_2	...	C_k	N

The following scores may be computed from the data in this contingency table:

2.1 Percent Hits *(PH)* (also called percent correct) is the percentage of time categorical hits occurred *(i=j),* considering all categories. It is a measure of accuracy and may also be referred to as the categorical percentage correct.

$$PH = \frac{\sum_{i=1}^{k} A_{ii}}{N} x100$$

2.2 Bias by Category (BIAS) measures the tendency to over-forecast (*BIAS* greater than 1) or under-forecast (*BIAS* less than 1) a particular category, *i,* of a multi-category contingency table (see Table A-1, where *k* values of bias exist.

$$BIAS_i = \frac{C_i}{R_i}$$

2.3 Probability of Detection (POD). A *POD* may be calculated for each individual category, *i,* of Table A-1. It measures the forecaster's success in covering each event of category *i* with a correct forecast, A_{ii}. The *POD* does not penalize the forecaster for incorrect forecasts of category

i.

$$POD_i = \frac{A_{ii}}{R_i}, \text{ where } i = 1, \ldots, k$$

Sometimes it is useful to combine two or more categories from a contingency table into a single category and compute a POD for the new category. For a description of this type of specialized contingency table and the POD formula, see sections 3 and 3.1.

2.4 False Alarm Ratio (FAR). A *FAR* may be calculated for each individual category, *i,* of Table A-1. It measures the fraction of forecasts of category *i* that were incorrect. It gets its name "false alarm" from the times when category *i* is a rare or extreme event that may require a warning, watch or advisory.

$$FAR_i = \frac{C_i - A_{ii}}{C_i}, \text{ where } i = 1, \ldots, k$$

Sometimes it is useful to combine two or more categories from a contingency table into a single category and compute an FAR for the new category. For a description of this type of specialized contingency table and the FAR formula, see sections 3 and 3.2.

2.5 Critical Success Index (CSI). A *CSI* may be calculated for each individual category, *i,* of Table A-1. It measures the forecaster's success in covering each event of category *i* with a correct forecast, A_{ii}, while also penalizing for incorrect forecasts of category *i.* It differs from the *POD* in that the *POD* doesn't penalize for incorrect forecasts.

$$CSI_i = \frac{A_{ii}}{R_i + C_i - A_{ii}}, \text{ where } i = 1, \ldots, k$$

Sometimes it is useful to combine two or more categories from a contingency table into a single category and compute a CSI for the new category. For a description of this type of specialized contingency table and the CSI formula, see sections 3 and 3.3.

2.6 Generalized Skill Score (SS). This generalized skill score measures the fraction of possible improvement of the forecasts over some standard or test set of forecasts.

$$SS = \frac{NC - E}{N - E}, \text{where:}$$

$$NC \text{ (number correct)} = \sum_{i=1}^{k} A_{ii}$$

and *E* represents some standard or test set of forecasts.

2.7 Heidke Skill Score (HSS). Sometimes the standard or test forecasts *(E)* from the generalized skill score (see section 2.6) are the values expected by chance and are computed from the marginal totals of the contingency table. One such score is the *HSS.*

$$HSS = \frac{NC - E}{N - E}, where:$$

$$NC\ (number\ correct) = \sum_{i=1}^{k} A_{ii}\ ; \quad E = \sum_{i=1}^{k} \frac{C_i R_i}{N}$$

A perfect Heidke skill score is one. Zero is indicative of no skill, and a negative score indicates skill worse than random forecasts. With three or more categories in the contingency table, Heidke only allows credit for categorical forecast hits along the diagonal of the contingency table, and therefore, does not penalize large categorical errors more than small categorical errors. This property rules out the possibility for granting "partial credit" to small forecast errors or "near hits." Also, correct forecasts of low frequency events are treated the same as correct forecasts of common events so the forecaster is not encouraged to forecast climatologically improbable (rare) events.

The CPC uses a version of the Heidke skill score for its main verification statistic. This is calculated by the formula:

$$HSS = \frac{NC - CH}{NT - CH} x100,$$

where NC is the total number of locations for which the forecast was correct, NT is the total number of locations for which a forecast was made, and CH is the number of locations which would be forecast correctly, on average, by chance. In a three class system (which is how all the CPC forecasts are characterized), one third of the locations are expected to be correct by chance. Thus if 99 locations are forecast, 33 are expected to be correctly forecast. This statistic results in scores of 100 if all locations are forecast correctly, zero if 33 are forecast correctly, and -50 if all locations are forecast incorrectly.

2.8 Peirce Skill Score *(PSS)*. The Pierce skill score (Peirce 1884), also known as the Hanssen–Kuipers discriminant (Hanssen and Kuipers, 1965) and the true skill statistic (Flueck 1987), is similar to the Heidke skill score. Peirce and Heidke differ only in how they estimate the number of correct forecasts that would be expected by chance in their respective denominators—the numerators of the two scores are identical. Both scores are equitable, which means that a perfect forecast (all correct) results in a score equal to one, and a no skill (random) forecast results in a score equal to zero. Negative scores are possible. With three or more categories in the contingency table, Peirce only allows credit for categorical forecast hits along the diagonal of the contingency table, and therefore, does not penalize large categorical errors

more than small ones. This property rules out the possibility for granting "partial credit" to small forecast errors or "near hits." Also, with three or more categories in the contingency table, correct forecasts of low frequency events are treated the same as correct forecasts of common events so the forecaster is not encouraged to forecast climatologically improbable events.

$$PSS = \frac{NC - E}{N - E^*}, where:$$

$$NC\ (number\ correct) = \sum_{i=1}^{k} A_{ii} \qquad E = \sum_{i=1}^{k} \frac{C_i R_i}{N} \qquad E^* = \sum_{i=1}^{k} \frac{R_i R_i}{N}$$

2.9 Equitable Skill Scores (ESS).

2.9.1 Subjective Explanation. Skill scores are often used to evaluate multi-category forecasts with a single score. Equitability is a desirable property for a skill score because equitability has the following characteristics:

a. A set of perfect forecasts (all categorical hits) produces a score equal to one.

b. A set of randomly generated forecasts or a set of forecasts that always predicts the same forecast category results in a "no skill" score equal to zero.

While equitable skill scores, such as Heidke (section 2.7) and Peirce (section 2.8), are convenient (they can often be computed by hand), they only grant credit for categorical forecast hits. Therefore, with three or more categories in the contingency table, Peirce and Heidke do not penalize large categorical errors more than small ones, and this rules out the possibility of receiving partial credit for "near hits." Also, correct forecasts of low frequency events are treated the same as correct forecasts of very common events so the forecaster is not encouraged to forecast climatologically improbable (rare) events.

Gandin and Murphy (1992) developed a mathematical framework for computing equitable scores that allow for a system of graduated, partial credit that considers the size of each miss and the observed frequency of each category. While Gandin and Murphy allowed for forecast systems with a higher number of forecast categories, examples of systems with greater than three categories were beyond the scope of their work. Gerrity (1992) built upon Gandin and Murphy and derived a general set of formulas that place no upper limit on the number of categories allowed in the system. Gerrity's formulas must be applied to scoring forecasts of ordinal variables (order matters) with maximum and minimum values, e.g., temperature, wind speed, ceiling, and visibility. While high speed computation is necessary for the Gerrity formulas, they are relatively simple to program. The Gerrity ESS has been implemented operationally in the NWS and has the following reward/penalty characteristics:

a. A relatively small reward is given for correctly forecasting common events.

b. A large reward is given for correctly forecasting rare events.

c. A graduated reward/penalty system is used, whereby a large forecast error for a given category is penalized more than a small forecast error for that category.

d. Less penalty is assigned to an incorrect forecast of a rare event than a similar size error of a common event. "Near hits" of rare events often receive a modest reward.

The otherwise favorable property of giving large rewards for correct forecasts of rare events may make the score volatile, especially with very small sample sizes of the rare events. For example, if a particular event occurs on a very rare basis, the ESS may increase substantially due to just one additional correct forecast of that rare event. Therefore, the ESS is not the ideal score for data requests that include relatively small geographic areas and/or short periods of time with little variability in the element. It is also important to exercise care in defining categories in the first place to keep very rare events and volatile scores from becoming a foregone conclusion.

Depending upon the element being verified, the rarest categories tend to be either the lowest or highest categories of the contingency table. For example with wind speed and significant wave height, the rarest events tend to be the highest categories. With ceiling and visibility, the rarest events tend to be the lowest categories. The ESS Low/High Category Delta is defined as the increase that occurs in the ESS due to one additional forecast hit in the lowest/highest category whose event count is at least one. Whenever the ESS is used, the delta values should always be checked for potential score volatility. A delta value that is unacceptably high should lead the user of *Stats on Demand* to resubmit a data request for a larger geographic area and/or longer time frame. See the last two paragraphs of section 2.9.2 for the mathematical definitions of the delta values.

2.9.2 Mathematical Background. The probability matrix, **P**, comes from the **A** matrix (Table A-1), where all

$$p_{ij} = \frac{A_{ij}}{N} \quad ; \quad (i = 1, \ldots, k \text{ and } j = 1, \ldots, k)$$

The row totals of the **P** matrix comprise **p**, the climatological probability vector, (p_1 , p_2 , ... , p_k). The column totals of the **P** matrix comprise **q**, the forecast probability vector, (q_1 , q_2 , ... , q_k).

Gandin and Murphy (1992) describe what is meant by an "equitable skill score" for the evaluation of categorical forecasts. The general formula is

$$ESS = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} s_{ij}$$

Note that p_{ij} are the elements in the aforementioned **P** matrix, and s_{ij} are the elements of the reward-penalty matrix, also called the scoring matrix (**S**). When an appropriate climatology is used to populate the **S** matrix, a random set of forecasts yields an ESS equal to zero, and a perfect set of forecasts (i.e., only the diagonal of the **P** matrix is populated) yields an ESS equal to one.

Gerrity (1992) derived the following formulas for populating the **S** matrix in a *k*-category system.
These formulas are only appropriate for ordinal variables (i.e., the order of the categories matters) that are not periodic. Wind speed and ceiling height are examples of ordinal, non-periodic variables. Wind direction is an example of an ordinal, periodic variable for which the Gerrity solution is not appropriate because as an eight-category variable, wind direction cannot "miss" by more than four categories (a non-periodic variable expressed in terms of eight categories can miss by up to seven categories).

Gerrity defines *p(r)* as the relative frequency with which category *r* of an event is observed in a large sample of forecasts and then defines *D(n)* and *R(n)*:

$$D(n) \equiv \frac{1 - \sum_{r=1}^{n} p(r)}{\sum_{r=1}^{n} p(r)} \qquad R(n) = \frac{1}{D(n)}$$

D(n) is the ratio of the probability that an observation falls into a category with an index greater than *n* to the probability that it falls into a category with an index less than or equal to *n*; *R(n)* is the reciprocal of this ratio of probabilities. In terms of *D* and *R*, Gerrity expresses the elements of a *k*-category equitable **S** matrix in the following manner:

$$s_{m,n} = \frac{1}{k-1}\left[\sum_{r=1}^{m-1} R(r) + \sum_{r=m}^{n-1}(-1) + \sum_{r=n}^{k-1} D(r)\right] \quad ; \quad n = (1,...,k)$$

$$s_{n,n} = \frac{1}{k-1}\left[\sum_{r=1}^{n-1} R(r) + \sum_{r=n}^{k-1} D(r)\right] \quad ; \quad 1 \le m < k, \quad m < n \le k$$

$$s_{n,m} = s_{m,n} \quad ; \quad 2 \le n \le k, \quad 1 \le m \le n$$

Burroughs (1993), appendix B, section n, applies these general equations for populating the **S** matrix to specific *k*-category marine elements.

The **S** matrix is computed directly from the sample of the *Stats on Demand* data request. This practice has one major shortcoming; requests for verification data from relatively small, restrictive samples will tend to produce volatile scores that fluctuate due to random changes in

the data set. Ironically, this problem is aggravated in these situations by the otherwise favorable ESS property of giving more weight to rare events. The following two paragraphs describe the measure used to help identify these situations.

Depending upon the element being verified, the rarest categories tend to be either the lowest or highest categories of the contingency table. To help the user of *Stats on Demand* test the ESS for volatility, one or both of the following "deltas" are calculated and listed in the verification reports with the ESS:

$$\delta_{low} = \frac{s_{aa}}{N}$$

$$\delta_{high} = \frac{s_{bb}}{N}$$

where δ_{low} is defined as the increase that occurs in the ESS due to one additional forecast hit in *a,* the lowest category in the contingency table whose total event count is at least one, and δ_{high} is defined as the increase that occurs in the ESS due to one additional forecast hit in *b,* the highest category in the contingency table whose total event count is at least one.

The user of *Stats on Demand* can easily calculate the delta for any intermediate category, *i*, in the contingency table by dividing the weight given in the reward-penalty matrix for a correct forecast in the *i*th category (s_{ii}) by the total sample size (N). The user of the ESS is strongly encouraged to pay close attention to the delta value provided with a particular score for an estimate of score volatility. If the score is too volatile for the user's tolerance, re-compute the score for a larger, less restrictive area in space and time.

3. Specialized Contingency Table. The following contingency table (Table A-2) may be used when only two outcomes (yes or no) exist for a given event or forecast, e.g., tornadoes. The number of correct forecasts for the specific event is given by *A*. The number of events observed but not forecast is given by *B*. The number of forecasts which did not verify is represented by *C*. The number of times the specific event was neither forecast nor observed is represented by *X.*

Table A-2 may be obtained from Table A-1 by combining multiple categories of Table A-1. For example with marine forecasts, sustained wind speeds are divided into seven categories. Define sustained wind speeds equaling or exceeding 28 knots (categories 6 and 7) as the "yes" outcome for a strong wind forecast or event. In this case, the "no" outcome is all sustained wind speeds less than 28 knots (categories 1 through 5 combined). The result is two categories (yes and no).

Table A-2. Specialized Contingency Table

		Forecasts	
		Yes	No
Events	Yes	*A*	*B*
	No	*C*	*X*

The scores most frequently computed from this table are:

3.1 Probability of Detection *(POD)* is the fraction of actual events *(A+B)* correctly forecast *(A)*. In the case of warnings, the *POD* is computed from the event database and is the number of warned events divided by the total number of events. The more often an event is correctly forecast, the better the score. The best possible score is 1, the worst possible score is 0.

$$POD = \frac{A}{A+B}$$

If *(A+B)* is the total number of events, e.g. turbulence or icing, sometimes it is useful to compute the *POD* of null events, e.g., no turbulence or no icing. Thus the *POD* of null events *(POD[N])* is the probability of null events that were forecast correctly. An alternative name for this statistic is the probability of null events *(PON)*. The formula is

$$POD[N] = PON = \frac{X}{X+C}$$

3.2 False Alarm Ratio *(FAR)* is the fraction of all forecasts *(A+C)* which were incorrect *(C)*. In the case of warnings, the *FAR* is computed from the event database and is the number of false alarms (unverified warnings) divided by the total number of warnings. The more often an event is forecast and does not occur, the worse the score. The best possible score is 0, the worst possible score is 1.

$$FAR = \frac{C}{A+C}$$

The *POD* and *FAR* are most often used in the verification of watches and warnings. However, it is possible to apply the *POD* and *FAR* to many events and forecasts related to public and aviation elements. Two examples are the *POD* for ceilings below 1000 feet and the *FAR* for forecasts of freezing rain.

Over-forecasting an event will achieve a high *POD* but at the expense of a high *FAR*. Overall

success can be expressed by the critical success index *(CSI).*

3.3 Critical Success Index is the ratio of correct forecasts *(A)* to the number of events *(A+B)* plus the number of incorrect forecasts *(C)*.

$$CSI = \frac{A}{A + B + C}$$

The best possible score is 1, the worst is 0. The relationship among *POD*, *FAR*, and *CSI* can be expressed as follows:

$$CSI = [(POD)^{-1} + (1 - FAR)^{-1} - 1]^{-1}$$

In the case of severe thunderstorm watches and warnings, the value of *A* varies depending upon whether it is taken from the warning or the event database. This is true because multiple events within a single county are sometimes counted as separate events in the event database, whereas only one warning can be in effect for a particular county at the same time. For this reason, the number of warned events in the event database, denoted below as A_e, may exceed the number of verified warnings in the warning database, denoted below as A_w. Using these conventions, the definitions of *POD* and *FAR* are

$$POD = \frac{A_e}{A_e + B}$$

$$FAR = \frac{C}{A_w + C}$$

Given these expressions for *POD* and *FAR* and the *CSI* formula, expressed in terms of *POD* and *FAR*, the *CSI* becomes:

$$CSI = \frac{A_w A_e}{A_w A_e + A_w B + A_e C}$$

4. Scores Computed for Specific Forecast Elements. Other scores may be computed, where N = number of cases; f_i = the *ith* forecast, and o_i = the *ith* observation (matching the forecast).

4.1 Temperature, Wind Speed and Direction, and Wave Height. Scores frequently computed for forecasts of temperature, wind speed and direction, and wave height include:

a. Mean Error *(ME)* indicates whether collective forecast values were too high or too low. This is also called the mean algebraic error.

$$ME = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)$$

b. Mean Absolute Error *(MAE)* measures error without regard to the sign (whether positive or negative).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\, f_i - o_i \,|$$

c. Root Mean Square Error *(RMSE)* weights large errors more than the MAE.

$$RMSE = \sqrt{\frac{1}{N}[\sum_{i=1}^{N}(fi - oi)^2]}$$

d. Measuring Errors Against Some Standard. The above measures of accuracy *(ME, MAE, RMSE)* may also be computed for some forecast standard, such as Model Output Statistics *(MOS)* guidance, climatology *(CLI)*, or persistence *(PER)*. For example, the *MAE* for *MOS* guidance forecasts (m_i) is

$$MAE_{MOS} = \frac{1}{N}\sum_{i=1}^{N}|\, m_i - o_i \,|$$

Forecast skill is determined by measuring the improvement of forecasts over a forecast standard. For example, the *MAE* may be used to compute the percent improvement of forecasts over *MOS*, $I(MAE)_{MOS}$.

$$I(MAE)_{MOS} = \frac{MAE_{MOS} - MAE}{MAE_{MOS}} x100$$

Other examples include $I(RMSE)_{MOS}$, $I(MAE)_{CLI}$, and $I(RMSE)_{PER}$.

4.2 Probability of Precipitation. Scores typically computed for probability of precipitation verification include:

a. Brier Score *(BS)* measures the mean square error of all PoP intervals forecast. The standard NWS Brier score, defined below, is one-half the original score defined by Brier (1950).

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

where, f_i = forecast probability for the *ith* case, o_i = observed precipitation

occurrence (0 or 1), and N = the number of cases.

b. Climatological Brier Score *(BS_{CLI})* is an application of the Brier score to forecasts, c_i, consisting of climatic relative frequencies, *RF* (see below).

$$BS_{CLI} = \frac{1}{N}\sum_{i=1}^{N}(c_i - o_i)^2$$

c. Improvement over Climate Based on Brier Score *($I(BS)_{CLI}$)* measures the improvement gained from actual forecasts versus climatological values.

$$I(BS)_{CLI} = \frac{BS_{CLI} - BS}{BS_{CLI}} x100$$

d. MOS Brier Score *(BS_{MOS})* is analogous to BS_{CLI}, except the Brier score is computed for MOS forecasts.

$$BS_{MOS} = \frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2$$

where, m_i = MOS guidance probability for the *ith* case. MOS guidance probabilities *(m_i)* are forecast to the nearest 0.01; however for NWS PoP verification, the m_i values are rounded to one of the following values: 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

e. Improvement over MOS Based on Brier Score *($I(BS)_{MOS}$)* is analogous to $I(BS)_{CLI}$, except this score measures the improvement of the forecast over MOS.

$$I(BS)_{MOS} = \frac{BS_{MOS} - BS}{BS_{MOS}} x100$$

f. Relative Frequency of an Event *(RF)* is the fraction of the time an event occurred.

$$RF = \frac{1}{N}\sum_{i=1}^{N} o_i$$

g. Reliability, a measure of bias, compares the number of forecasts of an event with the observed relative frequency of the event. The reliability may be determined overall or by forecast interval, e.g., 10 percent PoP intervals or (0, 5, 10, 20, 30, . . ., 80, 90, 100.

$$\frac{1}{N}\sum_{i=1}^{N} f_i \quad \text{compared with} \quad \frac{1}{N}\sum_{i=1}^{N} o_i \quad ,$$

where, N is the total number of events or the number of events in the interval. If the number of forecasts of the event or interval is larger (smaller) than the observed relative frequency of the event or interval, the event or interval was overforecast (underforecast).

4.3. QPF.

a. Bias, Threat Score, POD, and FAR, when applied to QPF verification, are computed from gridded data for specific precipitation amount thresholds, e.g. 0.01 inch, 0.25 inch, 0.50 inch, 1.00 inch, etc. Bias *(B)* and Threat Score *(TS)* (Gilbert 1884; Junker et al. 1989; Schaefer 1990) (also known as the *CSI*) are defined as follows:

$$B = \frac{F}{O}$$

$$TS = CSI = \frac{H}{F + O - H}$$

where, F is the number of points forecast to have at least a certain amount (threshold) of precipitation, O is the number of points observed to have at least the threshold amount, and H is the number of points with correct forecasts for that threshold of precipitation. When the bias is less [greater] than unity for a given threshold, the forecast is under [over] forecasting the areal coverage for that amount.

Geometrically, the threat score for a given threshold amount represents the ratio of the correctly predicted area to the threat area. Threat area is defined as the envelope of forecast and observed areas for that threshold. A perfect forecast yields a threat score of one, and a forecast with no areas correctly predicted receives a zero. The threat score, therefore, provides a measure of how accurately the location of precipitation is forecast within the valid period of the forecast. To receive a high threat score, forecast precipitation must be accurate—both spatially and temporally. For example, if a 1.00-inch isohyet is forecast, and all the observed rainfall within that area ranges from 0.8 to 0.99 inch, the forecaster's 1.00-inch threat score would be zero. However, the 0.8 to 0.99 inch area would favorably affect the 0.5-inch threat score. Also, a forecast area that is adjacent to an observed area with no overlap produces a zero threat score, and forecasts that are incorrect by just a couple of hours may receive little or no credit. Closely related to the threat score are *POD* and *FAR* which are expressed as:

$$POD = \frac{H}{O}$$

$$FAR = \frac{F - H}{F}$$

b. Equitable threat score *(ETS)* (Messinger 1996) is similar to the threat score except the expected number of hits in a random forecast, *E,* is subtracted from the numerator and denominator:

$$ETS = \frac{H - E}{F + O - H - E}$$

where $E=FO/N,$ and N is the number of points verified. E is substantial for low precipitation categories, i.e., 0.10 inch or less in 24 hours, small at intermediate categories, and negligible for high categories, i.e., 1 inch or more in 24 hours.

4.4 Ceiling Height and Visibility. The Log Score *(LS)* is used for verifying ceiling height and visibility forecasts. It emphasizes accuracy in the more critical lower ceiling height and visibility ranges.

$$LS = \frac{50}{N}\sum_{i=1}^{N} \left| Log_{10}\left(\frac{f_i}{o_i}\right)\right|$$

Where f_i is the category of the *ith* forecast and o_i is the category of the *ith* observation. Note, f_i and o_i may also be used to represent the actual respective forecast and observed values of the element (i.e., ceiling height in feet, visibility in statute miles). Persistence is often used as the reference standard for evaluating ceiling height and visibility forecasts. The last hourly observation available to the forecaster before dissemination of the terminal aerodrome forecast defines the persistence forecasts of ceiling height and visibility to which the TAFs are compared.

4.5 Aviation Weather Center (AWC) Verification Statistics. The following statistics are used for verifying AWC forecasts:

a Probability of Detection *(POD)*. Same as section 3a of this appendix.

b. False Alarm Ratio *(FAR)*. Same as section 3b of this appendix.

c. Probability of Detection of "No" Observations *(POD[N])* is an estimate of the proportion of "no" observations that were correctly forecast (i.e., PIREPs which include reports such as negative icing or negative turbulence). An alternative name for this statistic is the probability of null events *(PON)*. Based on the

contingency table presented in section 3 of this manual,

$$POD[N] = PON = \frac{X}{X + C}$$

d. Percent Area *(% Area)* is the percentage of the forecast domain's area where the forecast variable is expected to occur. It is the percent of the total area with a YES forecast.

e. Percent Volume *(% Vol)* is the percentage of the forecast domain's volume where the forecast variable is expected to occur. It is the percent of the total volume with a YES forecast.

5. References.

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. Monthly Weather Review, **78**, 1-3.

Burroughs, L.D., 1993: National marine verification program - verification statistics. OPC Technical Note/NMC Office Note No. 400, National Weather Service, NOAA, U.S. Dept. of Commerce, 48 pp.

Burroughs, L.D., 2002: Verification scores from performance matrices—a short tutorial. Personal communication, NOAA, National Weather Service, National Centers for Environmental Prediction, Environmental Modeling Center, Ocean Modeling Branch (W/NP21).

Flueck, J.A., 1987: A study of some measures of forecast verification. *Preprints, 10th Conference on Probability and Statistics in the Atmospheric Sciences.* Edmonton, AB, Canada, American Meteorological Society.

Gandin, L.S., and A.H. Murphy, 1992: Equitable skill scores for categorical forecasts. Monthly Weather Review, **120**, 361-370.

Gerrity, J.P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.,* **120,** 2709-2712.

Gilbert, G.F., 1884: Finley's tornado predictions. American Meteorological Journal, **1**, 166-172.

Hanssen, A.W. and W.J.A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Mededeelingen en Verhandelingen,* Royal Netherlands Meteorological Institute, **81.**

Hughes, L.A., 1980: Probability forecasting - reasons, procedures, problems. NOAA

Technical Memorandum NWS FCST 24, National Weather Service, NOAA, U.S. Department of Commerce, 84 pp.

Junker, N.W., J.E. Hoke, and R.H. Grumm, 1989: Performance of NMC's regional models. Weather and Forecasting, 4, 368-390.

Livezey, R.E., 2003: Categorical events (chapter 4). *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Edited by I.T. Jolliffe and D.B. Stephenson, John Wiley and Sons, Ltd., 240 pp.

Messinger, F., 1996: Improvements in precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. Bulletin of the American Meteorological Society, 77, 2637-2649.

Peirce, C.S., 1884: The numerical measure of the success of predictions. *Science,* **4,** 453-454.

Schaefer, J.T., 1990: The critical success index as an indicator of warning skill. Weather and Forecasting, **5**, 570-575.

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