1.2 TRANSFORMED SKILL SCORES

Tressa L. Fowler, Randy Bullock, and Barbara G. Brown
National Center for Atmospheric Research, Boulder, Colorado

1. INTRODUCTION

Skill scores are commonly used in forecast verification as a means of comparing forecasts. Because skill scores attempt to quantify a multidimensional problem in a single dimension, skill scores are inherently problematic. They are, however, unlikely to be dropped from use. Skill scores are similar in many ways to Goodness-of-Fit (GOF) tests. In particular, observed discrete multinomial data can be compared to a model via the GOF tests. Skill scores compare categorical forecasts to baseline forecasts such as chance, persistence or climatology, or the current operational standard.

Read and Cressie (1988), hereafter RC88, have derived a generalized form for GOF tests. This generalized form, called the power divergence family of statistics, requires a parameter. This parameter is most importantly the exponent of the ratio of counts. The work on GOF tests provides motivation for transformations applied to skill scores, as described in section 2.1. Section 2.2 presents the three methods of transformation that are applied to the skill scores. The effect of exponentiation on skill scores derived from the 2x2 contingency table is investigated for two sets of forecasts, presented in section 3. The skill scores of interest are listed in section 4. The results of the empirical investigation of the effects of the transformations on the data sets are presented in section 5. Finally, section 6 offers conclusions and describes future work.

2. MOTIVATION AND METHODS

2.1 Motivation

It is well documented that all skill scores suffer from some undesirable qualities (e.g. Marzban, 1998). The same is true of GOF tests. In fact, Cressie and Read (1984) say of GOF tests that there "is no uniformly preferable test". However, they also demonstrate that use of different exponents can mitigate some of the problems associated with different tests.

GOF tests, such as Pearson's Chi-squared and Freedman-Tukey, are generally used to explicitly test the hypothesis that multinomial data come from some specified distribution. In order to conduct this test, the distribution of the GOF statistic must be known, at least asymptotically. Skill scores, such as the Heidke Skill Score (HSS) and True Skill Statistic (TSS), are used in much the same way, to "test" if a forecast is "better" than some standard. However, this test is implicit. The achieved score is not compared to some distribution or standard to determine the outcome. Indeed, this would be impossible, as the distributions of the various skill scores are not even known.

According to RC88, the general format for the Power Divergence family of statistics is:

$$PD(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_i O_i \left[\left(\frac{O_i}{E_i} \right)^{\lambda} - 1 \right] \qquad (1)$$

The usual GOF tests are achieved by setting the parameter λ to the appropriate value (e.g. for Pearson's χ^2, $\lambda = 1$; for the Likelihood ratio statistic G^2, take the limit of PD as $\lambda \rightarrow 0$). The power divergence statistic seems to work well in a variety of situations when $\lambda = 2/3$, as it provides a compromise between the most well known tests (RC88). For instance, the test maintains a balance between determining lack of fit yet remaining robust to a single cell departure. The test with this parameter also works in situations of sparse cell counts or finite populations. Additionally, the moments very closely match the asymptotically derived Chi-square moments.

The O_i (E_i) are the observed (expected) counts for the i^{th} cell. The exponent λ is applied to the ratio of the observed counts to expected counts. The additional factor of O_i can be interpreted as a weight for the ratio (RC88).

The general format of skill scores is a ratio of differences (Stanski et al, 1989; Wilks, 1995). The numerator measures the difference between the forecast of interest and some reference forecast while the denominator measures the difference between a perfect forecast and the reference forecast. This formula involves the measure of the forecast to be evaluated (A), the measure of some reference forecast (A_{ref}) and the measure of a perfect forecast (A_{perf}). These measures are combined as a ratio of differences, with the following formula:

$$\frac{A - A_{ref}}{A_{perf} - A_{ref}} \qquad (2)$$

The result is interpretable as a "percentage improvement" over the reference forecast (Wilks, 1995). However, comparing different skill scores for the same forecasts makes it clear that the "percent improvement" varies considerably depending on the score chosen.

While GOF tests are equivalent except for the choice of exponent, this is not the case with skill scores. Instead, they differ in the quantities used to measure the forecast (A) and reference forecast (A_{ref}).

Corresponding author address: Tressa L. Fowler, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307. e-mail: tressa@ucar.edu

2.2 Methods

GOF tests involve only two sets of counts. Skill scores incorporate a third set of counts, those of the reference forecast. Therefore, adding an exponent to the ratio of two counts is not practical with skill scores. However, the numerator and denominator of the skill score are just adjusted counts, so the exponent can be applied to the entire skill score (SS). This method of transformation, SS^{λ}, is hereafter referred to as method one.

Power transformations, including the natural logarithm, are useful for count data (Hoaglin et al, 1983). Generally, counts and amounts are skewed, bounded below, and have long right tails. Additionally, the variability frequently increases as the count or amount increases (i.e. the data are heteroscedastic). Application of a power transformation can yield a more gaussian looking sample; one that is symmetric and/or homoscedastic.

A second method of transforming the skill scores is achieved by applying power transformations to the counts prior to computing the skill score. The square root and natural logarithm are most commonly used transformations for counts. Here we will investigate the properties of power transformations with exponents ranging from (*0*, *2]* and the natural logarithm.

A third proposed method involves computing:

$$\left(\frac{A}{A_{perf} - A_{ref}} \right)^{\lambda} - \left(\frac{A_{ref}}{A_{perf} - A_{ref}} \right)^{\lambda} \quad (3)$$

For all methods, when $\lambda = 1$, the standard skill score is obtained. When the original skill score is 0, the transformation has no effect, so a no skill forecast is mapped to a no skill forecast by the transformation. These transformations apply only to positive skill scores, as taking non-integer powers of negative numbers can yield complex numbers, e.g. $\sqrt{-1}$. If transformations for negative scores were of interest, the transformations could be determined for the absolute values, with the negative sign replaced at the end.

3. DATA

Skill scores are computed on two sets of forecasts described in this section. Presented in Section 3.1 are the infamous Finley tornado forecasts that began the whole skill score debate over a century ago. Some model output statistics (MOS) precipitation forecasts are detailed in section 3.2.

3.1. The Finley Tornado Forecast Data

Three versions of the Finley tornado forecasts are considered, as reproduced from Stephenson (2000). The first table, 1a, contains the original Finley tornado forecasts. Table 1b has same forecasts "hedged" to achieve unbiased forecasts. Finally, table 1c contains random forecasts with the same marginal totals as the Finley forecasts.

In each of these tables, the counts in a single cell dwarf the other counts. This is characteristic of forecasts of "rare events". This type of table causes much trouble for many skill scores. Analogously, some GOF tests behave poorly when the cell probabilities are unequal (Koehler and Larntz, 1980).

Table 1a: Contingency table of Finley tornado forecasts.

Forecast	Observed		
	YES	NO	Total
YES	28	72	100
NO	23	2680	2703
Total	51	2752	2803

Table 1b: Contingency table of unbiased or "hedged" Finley tornado forecasts.

Forecast	Observed		
	YES	NO	Total
YES	14	37	51
NO	37	2715	2752
Total	51	2752	2803

Table 1c: Contingency table of random forecasts with same marginal totals as Finley tornado forecasts.

Forecast	Observed		
	YES	NO	Total
YES	2	98	100
NO	49	2654	2703
Total	51	2752	2803

3.2. The Precipitation Data

Tables 2a, b, and c present MOS precipitation forecasts as collapsed into 2x2 contingency tables. The original data are from Goldsmith (1990), but these contingency tables are reproduced from Wilks (1995).

Table 2a: Contingency table of MOS Freezing Rain forecasts.

Forecast	Observed		
	YES	NO	Total
YES	50	162	212
NO	101	6027	6128
Total	151	6189	6340

Table 2b: Contingency table of MOS Snow forecasts.

Forecast	Observed		
	YES	NO	Total
YES	2364	217	2581
NO	296	3463	3759

Total	2660	3680	6340

Table 2c: Contingency table of MOS Rain forecasts.

Forecast	Observed		
	YES	NO	Total
YES	3288	259	3547
NO	241	2552	2793
Total	3529	2811	6340

4. SCORES

Four skill scores are used in this analysis. They include the HSS, the TSS, the Gilbert Skill Score (GSS), and the Probability of Detection Skill Score (PODSS). Each of the four skill scores is computed from the cell counts in a contingency table. While several other measures can be computed from the contingency table counts, most do not follow the format of skill scores (e.g. CSI and FAR). They are therefore excluded from these analyses.

Table 3: Contingency Table values used to compute skill scores.

Forecast	Observed		
	YES	NO	Total
YES	a	b	a+b
NO	c	d	c+d
Total	a+c	b+c	n

The formulas for the skill scores, expressed in terms of the cell counts in Table 3, are listed in equations 4 through 7 below. HSS is defined as

$$HSS = \frac{(a+d) - ((a+b)(a+c)+(b+d)(c+d))/n}{n - ((a+b)(a+c)+(b+d)(c+d))/n} \quad (4)$$

The TSS does not precisely follow the formula for skill scores. Instead, the measure of reference used in the denominator is constrained to be unbiased (A_{ref}´). When the forecast to be evaluated is unbiased, the TSS is equivalent to the HSS.

$$TSS = \frac{(a+d) - ((a+b)(a+c)+(b+d)(c+d))/n}{n - \left((a+c)^2 + (b+d)^2\right)/n} \quad (5)$$

The GSS ignores the cell count *d*, in an effort to prevent the score from being swamped by the very common non-rare event.

$$GSS = \frac{a - (a+b)(a+c)/n}{(a+b+c) - (a+b)(a+c)/n} \quad (6)$$

Finally, the PODSS is the standard Probability of Detection (POD) corrected by the number of correct forecasts expected based on chance (Schaeffer, 1990). It differs from (4) only by the subtraction of b from the denominator.

$$PODSS = \frac{a - (a+b)(a+c)/n}{(a+c) - (a+b)(a+c)/n} \quad (7)$$

5. RESULTS

For each of the transformation methods described in Section 5, the scores are compared to the other skill scores computed for the same data and to the same skill score computed for the hedged and random data.

5.1 Method One

Method one is the simplest transformation of the skill scores as it consists solely of exponentiating the score. Figure 1 shows a graph of the GSS for the precipitation forecast data. The scores decrease from 1 to 0 as the exponent increases.

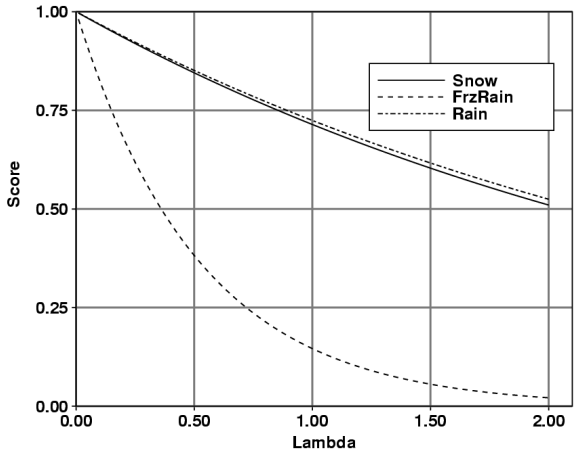

Figure 1: GSS transformed by method one for the precipitation forecast data.

Graphs of the other scores and other data sets look similar, and have been excluded for the sake of brevity. This method does not seem provide any benefit over the original skill score. The exponent could, in theory, be chosen to make the score appear larger or to maximize the difference between the scores of different forecasts.

5.2 Method Two

Method two consists of applying power transformations to the contingency table counts before computing the skill score. Figure 2 shows a graph of all the scores plus the POD for the original Finley data (F), the square root of the Finely data (R), and the natural logarithm of the Finley data (L).

For the Finley data, the transformations have little effect on the POD. However, the values of the different skill scores are closer (i.e. more consistent) when the transformations have been applied. It is important to keep in mind that the Finley forecasts are for a “rare event”.

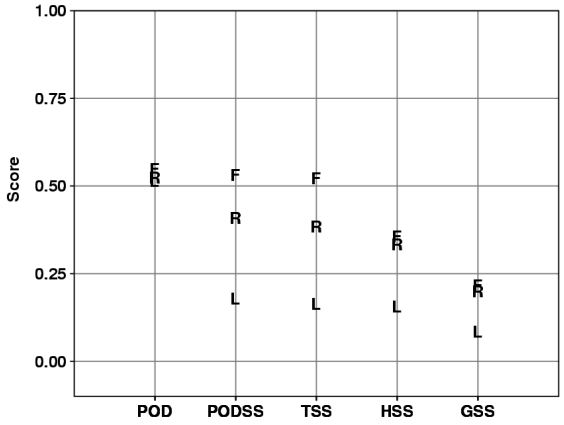

Figure 2: Skill scores for Finley's tornado forecasts (F), their square root (R), and natural logarithm (L).

In Figure 3, the effects of bias on the scores and transformed scores are illustrated. For all scores, the square root of the Finley forecast data (R) is closer to the score for the unbiased forecasts (U) than are the original scores (F). Taking the square root of the unbiased forecasts (S), however, has little effect, as these values lie very close to the unbiased forecasts (U).

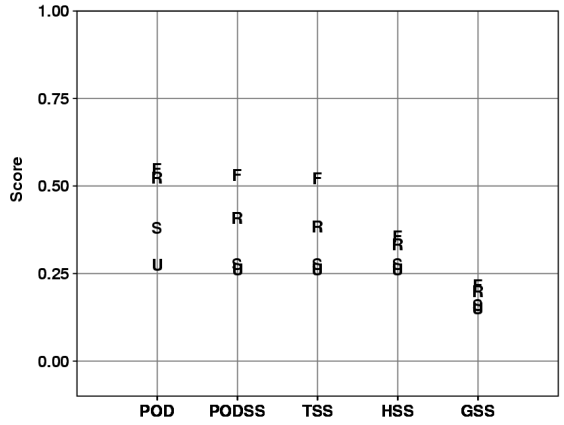

Figure 3: Skill scores for Finley's original tornado forecasts (F), the unbiased version of Finley's forecasts (U), and their respective square roots (R) and (S).

Figure 4 shows the same graph for the snow data (S), along with its square root (R) and natural logarithm (L). For the snow data, the transformations have a bigger effect on all of the measures. For this data, the skill scores on the untransformed data were already fairly consistent, therefore the transformations do not seem to yield more consistent scores. The transformation also has the effect of lowering the scores. In practice, this transformation should probably include some constant term to adjust the result back to an appropriate level, serving the same purpose as the $2\lambda/(\lambda+1)$ term in the power divergence statistic.

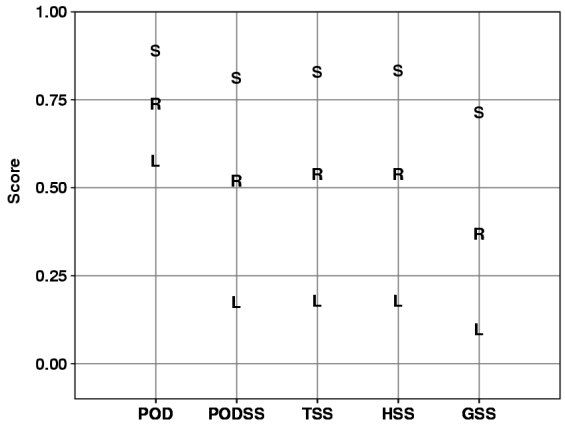

Figure 4: Skill scores for snow forecasts (S), their square root (R) and natural logarithm (L).

Figure 5 shows the effect of transforming the cell counts of the original Finley tornado forecasts on each of the scores. Figure 6 shows the effect of transforming the cell counts of the snow forecasts on each of the scores. Clearly, the effect of the transformation depends very much on the original counts. For the Finley data, the transformations yield small changes in the scores for exponents in the interval [0.5, 1.5]. However, for the snow forecasts, the change is much greater for the exponent in the same interval.

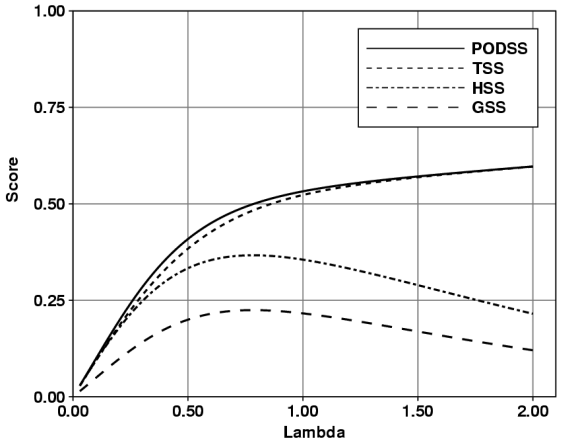

Figure 5: Method 2 transformed skill scores for original Finley tornado forecasts.

This transformation, with $\lambda < 1$, "brings in" the larger counts. This reduces the effects of rare events and bias on the scores. This transformation is monotone in some cases but not in others. This can be remedied by restriction of λ to some interval around one.

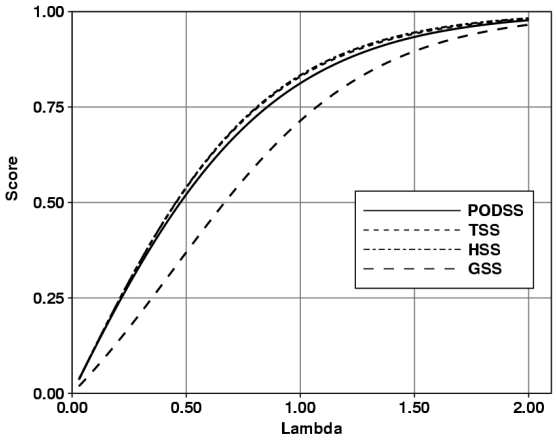

Figure 6: Same as figure 5 for snow forecasts.

5.3 Method Three

Method three applies equation (3) to the skill scores. When $\lambda > 1$, the skill scores can quickly get very large. However, for $\lambda < 1$, the differences in the scores are smaller. Figures 7 through 10 show the each of the four skill scores transformed by method 3 on the three versions of the Finley data.

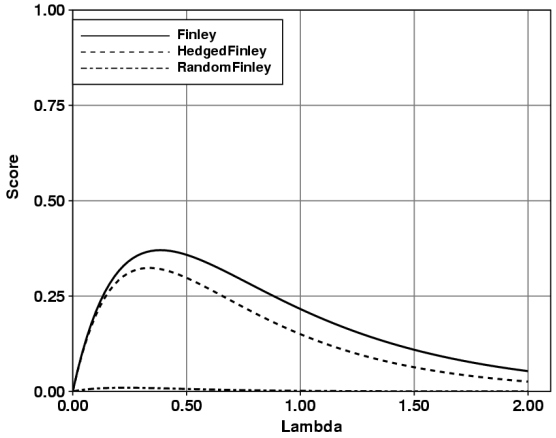

Figure 7: Method 3 transformed GSS for all versions of the Finley data.

The GSS and PODSS graphs look fairly similar to each other, as do the HSS and TSS graphs. However, the GSS and PODSS are very different from the HSS and TSS graphs. The first two seem to maintain a roughly equal distance between the scores for the original and hedged forecasts in the interval [0.6, 1.5]. Additionally, these transformed scores are monotone decreasing on this interval. The HSS and TSS are near zero for $\lambda > 0.5$, but increase dramatically when $\lambda > 1$. For the first two scores, GSS and PODSS this transformation is well behaved if restricted to some interval around one. However for the HSS and TSS, a very small change in λ results in a very large change in the score.

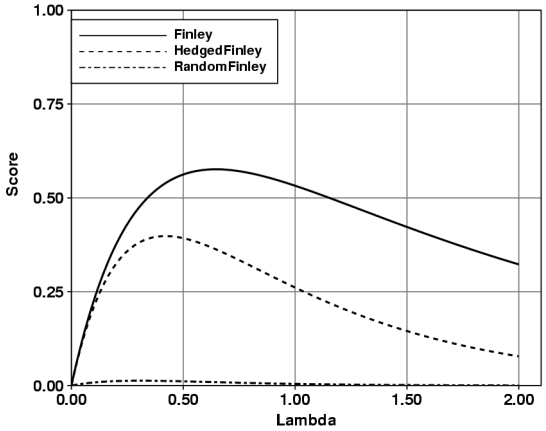

Figure 8: Method 3 transformed PODSS for all versions of the Finley data.

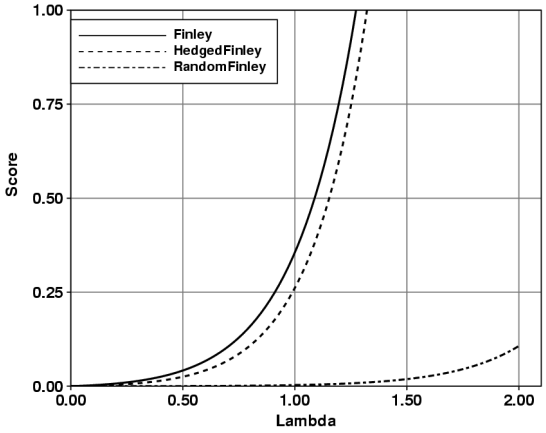

Figure 9: Method 3 transformed HSS for all versions of the Finley data.

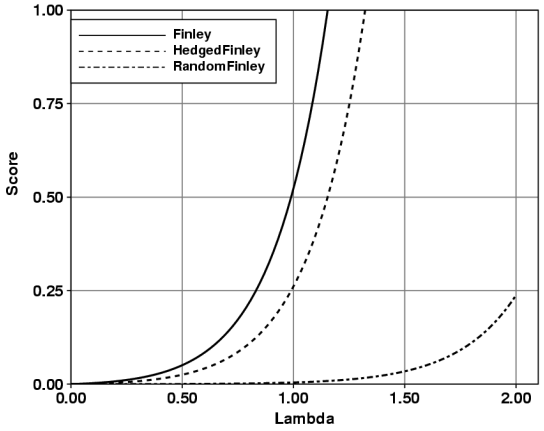

Figure 10: Method 3 transformed TSS for all versions of the Finley data.

6. CONCLUSIONS AND FUTURE WORK

As with GOF tests, no skill score is preferable in all situations. Both GOF tests and skill scores can exhibit undesirable behavior when cell counts are too small, too large, or too different from each other. In the forecast verification situation, these types of cell counts are the only types of interest. (Equal cell counts would correspond to "coin flip" forecasts paired with an event that has a 50% climatological probability.)

Restriction of λ to some interval around one is necessary for all of the types of transformations attempted here. Some of the transformed scores are not monotonic near $\lambda = 0.5$. Others increase steeply when $\lambda > 1$.

Method two has some nice properties when $\lambda < 1$. For example, the cell counts are brought closer to equal and the effect of bias o the score is reduced. For some of the forecasts, the power transformations have no effect on the resulting score. This method seems to yield more consistent results between the different scores in some cases. Previous research supports use of power transformations on count data.

Method three is the most complicated of the three transformations. It does not seem to yield more consistent results between the scores. For some forecasts, the transformation yields extremely high values for some scores and very low values for other scores. This transformation has a very different effect on the GSS and PODSS than it does on the HSS and TSS.

For this study, all samples were large. The effect of transformations on smaller sample sizes should be assessed. A more mathematically rigorous examination of the properties of the various transformations on skill scores will be attempted. Additionally, the effect of using weights on ratios of counts should be investigated.

ACKNOWLEDGEMENTS

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

REFERENCES

Cressie, N., and T. R. C. Read, 1984: Multinomial Goodness-of-Fit Tests, *J. R. Statist. Soc.* B, **46,** 440-464.

Goldsmith, B. S., 1990: *NWS Verification of Precipitation Type and Snow Amount Forecasts During the AFOS Era*, NOAA Tech. Mem. NWS FCST 33. National Weather Service, Camp Springs, MD, 28pp.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey, 1983: Understanding Robust and Exploratory Data Analysis. John Wiley and Sons, Inc., New York.

Koehler, K. J., and K. Larntz, 1980: An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials. *JASA*, **75,** 336-344.

Marzban, C., 1998: Scalar Measures of Performance in Rare-Event Situations. *Weather and Forecasting*, **13**, 753–763.

Read, T. R. C. and N. A. C. Cressie, 1988: Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer Verlag, New York.

Schaeffer, J. T, 1990: The Critical Success Index as an Indicator of Warning Skill, *Weather and Forecasting,* **5,** 570-575.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of Common Verification Methods in Meteorology. Atmospheric Environment Service, Forecast Research Division, Ontario, Canada.

Stephenson, D. B, 2000: Use of the "Odds Ratio" for Diagnosing Forecast Skill, *Weather and Forecasting,* **15,** 221-232.

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