NOTES AND CORRESPONDENCE

Visualizing Multiple Measures of Forecast Quality

PAUL J. ROEBBER

Atmospheric Science Group, Department of Mathematical Sciences, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin

(Manuscript received 19 May 2008, in final form 24 July 2008)

ABSTRACT

A method for visually representing multiple measures of dichotomous (yes–no) forecast quality (probability of detection, false alarm ratio, bias, and critical success index) in a single diagram is presented. Illustration of the method is provided using performance statistics from two previously published forecast verification studies (snowfall density and convective initiation) and a verification of several new forecast datasets: Storm Prediction Center forecasts of severe storms (nontornadic and tornadic), Hydrometeorological Prediction Center forecasts of heavy precipitation (greater than 12.5 mm in a 6-h period), National Weather Service Forecast Office terminal aviation forecasts (ceiling and visibility), and medium-range ensemble forecasts of 500-hPa height anomalies. The use of such verification metrics in concert with more detailed investigations to advance forecasting is briefly discussed.

1. Introduction

Demand for forecasts occurs in any endeavor for which the future is of interest and is unknown, thus extending beyond the natural sciences to business, economics, transportation, and sports (e.g., Jolliffe and Stephenson 2003). Even where certainty exists, as with Benjamin Franklin's oft-quoted observation concerning death and taxes, the details remain of interest. Hence, the existence of actuarial tables and retirement planning, both of which can be considered forms of ''climatology.'' In the domain of weather sensitive human activities, the need for meteorological forecasts has motivated progress in meteorological science since the nineteenth century (e.g., Gribbin and Gribbin 2004; Willis and Hooke 2006).

Obstacles to meteorological forecast progress arise in three categories: scientific, such as incomplete understanding of atmospheric phenomena and their intrinsic chaotic nature; technical, as with limits to computer power and the resolution of a specific observing platform; and political, for example, the sparseness of observations at certain temporal–spatial scales and the structures of forecast services themselves. Systematic forecast verification can assist in overcoming such obstacles by allowing the quality of the forecasts to be assessed and, through appropriate feedback, to revise and hopefully improve procedures. In particular, Ramage (1993) notes that the posing and testing of hypotheses is a natural activity in a forecast setting. As verification results come in, hypotheses can be revised or discarded based on the new information.

It is here, however, that difficulties emerge. As has been amply discussed in the literature, forecast verification is a complex and multidimensional problem (e.g., Murphy 1993; Brooks and Doswell 1996; Katz and Murphy 1997; Wilks 2006; Morss et al. 2008 and many others). Distribution-oriented verification approaches (Murphy and Winkler 1987; Murphy et al. 1989; Brooks and Doswell 1996), which retain the relationships between forecasts and observations, are more comprehensive than measures-oriented methods in representing the many aspects of forecast quality (Murphy 1993). Despite this, summary measures, although necessarily incomplete, may be convenient in many circumstances and remain in frequent use (Jolliffe and Stephenson 2003). It is clear that summary measures of forecast quality are the

Corresponding author address: Prof. Paul J. Roebber, Atmospheric Science Group, Dept. of Mathematical Sciences, University of Wisconsin—Milwaukee, 3200 North Cramer Ave., Milwaukee, WI 53211.
E-mail: roebber@uwm.edu

DOI: 10.1175/2008WAF2222159.1

2x2 Contingency Table		Event Observed	
		Yes	No
Event Forecast	Yes	A	B
	No	C	D

FIG. 1. The 2 × 2 contingency table.

start rather than the endpoint of investigations intending to understand and improve performance.

In this study, a 2 × 2 contingency table is constructed from dichotomous (yes–no) forecasts and well-known quality measures are derived from the entries of this table (Fig. 1). In section 2, it is shown that several of these commonly used measures are mathematically related and can be geometrically represented in a single diagram. As such, the accuracy, bias, reliability, and skill (obtained by plotting a measure of accuracy relative to a reference forecast) can be simultaneously visualized and the underlying relationships easily understood. Results from two previously published verification datasets [snow density by Roebber et al. (2003) and convective initiation by Fowle and Roebber (2003)] and four new datasets (severe convection, heavy rainfall, terminal aviation forecasts, and medium-range forecasts of 500-hPa height anomalies) using these measures are presented in this manner in section 2. A summary is presented in section 3.

2. The geometric relationship between POD, FAR, CSI, and bias

Taylor (2001) exploited the geometric relationship between three measures of model performance—correlation, normalized root-mean-square difference, and variance—to represent them in a single diagram. For a good model, the correlation between forecasts and observations is high, the root-mean-square difference is small, and the variances are similar. The position of the model point in the diagram relative to the observed reference gives an immediate visualization of accuracy. Lambert and Boer (2001) independently obtain a similar result.

One can proceed in a similar fashion with dichotomous (yes–no) forecasts. The forecasts and corresponding observations in such a system result in a 2 × 2 contingency table (Fig. 1). Key quality measures in this system, the probability of detection (POD), false alarm ratio (FAR), bias, and critical success index (CSI; also known as the threat score), are then defined:

$$\mathrm{POD} = \frac{A}{A + C}, \tag{1}$$

$$\mathrm{FAR} = \frac{B}{A + B}, \tag{2}$$

$$\mathrm{bias} = \frac{A + B}{A + C}, \quad \text{and} \tag{3}$$

$$\mathrm{CSI} = \frac{A}{A + B + C}. \tag{4}$$

With some algebraic manipulation, POD, FAR or its equivalent, success ratio (SR = 1 − FAR), bias, and CSI can be related as follows:

$$\mathrm{CSI} = \frac{1}{\frac{1}{\mathrm{SR}} + \frac{1}{\mathrm{POD}} - 1} \quad \text{and} \tag{5}$$

$$\mathrm{bias} = \frac{\mathrm{POD}}{\mathrm{SR}} = \tan\theta, \tag{6}$$

where θ is the angle from the abscissa. Equation (5) was previously derived by Schaefer (1990).

Figure 2 is an example of a diagram so constructed. For good forecasts, POD, SR, bias, and CSI approach unity, such that a perfect forecast lies in the upper right of the diagram. Deviations in a particular direction will indicate the relative differences in POD and SR, and consequently bias and CSI. As with the Taylor (2001) diagram, an immediate visualization of differences in performance is thus obtained. Further, skill can be assessed by plotting the forecast quality measure relative to a reference forecast (climatology, persistence, or any other desired baseline).

It is important to note that for a 2 × 2 contingency table, there are three degrees of freedom. For example, Stephenson (2000) shows that the four elements of the 2 × 2 contingency table (*A*–*D* in Fig. 1) can be completely expressed by three distinct verification measures: the hit rate, false alarm rate, and bias. In this sense, the formulation expressed by Eqs. (1)–(6) and its related diagram is incomplete in that it neglects the number of times for which a null event is forecast and no event occurs (element *D* of the contingency table). In rare event forecasting, however, neglecting this term can be useful since skill can be inflated by verification of the abundant and often trivial nonevents. In particular, Doswell et al. (1990) show that two common measures of forecast quality, the true skill statistic (TSS) and the Heidke skill score (HSS), approach the POD and a simple function of the CSI, respectively, in the limit as element *D* becomes large. In instances where the forecast event of interest is less rare, however, it may be important to account for nonevents in the verification.

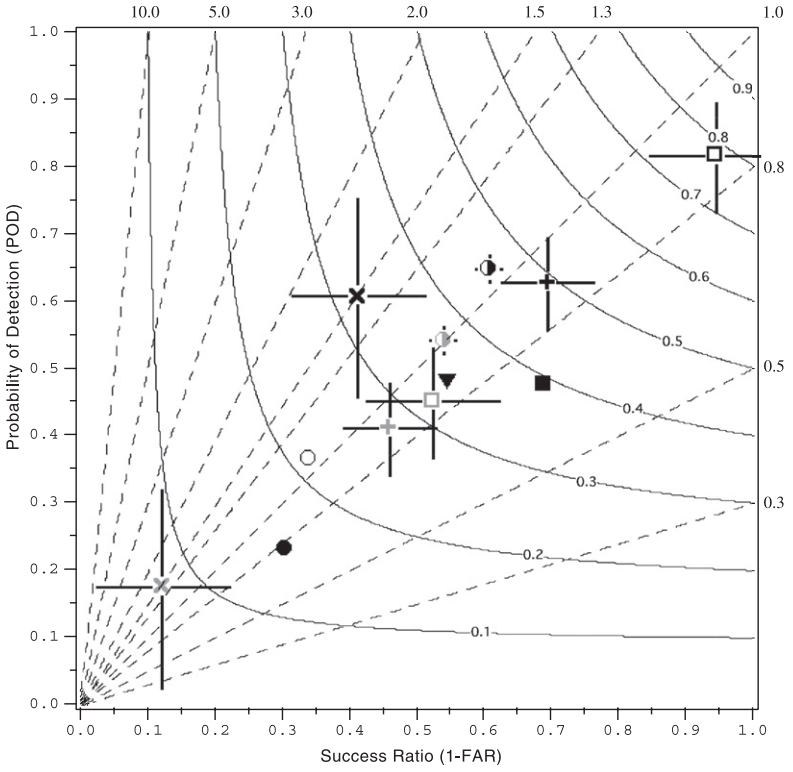

FIG. 2. Performance diagram summarizing the SR, POD, bias, and CSI. Dashed lines represent bias scores with labels on the outward extension of the line, while labeled solid contours are CSI. Sampling uncertainty is given by the crosshairs and reference ''sample frequency'' forecasts (where available) are in gray. Shown are 48-h forecasts of convective occurrence (bold square; Fowle and Roebber 2003), SPC forecasts of significant severe weather (filled square) and significant tornadoes (filled triangle), HPC 6–24-h forecasts of heavy precipitation (amounts in excess of 12.5 mm in 6 h) for cold (open circle) and warm (filled circle) seasons, heavy and light snow densities from a neural network [boldface multiplication and plus symbols, respectively; Roebber et al. (2003)], TAFs from NWS WFOs (bold half-filled circle), and MOS (gray half-filled circle).

One approach might be to augment the above verification with the percent correct rejections,

$$\mathrm{PCR} = \frac{D}{B + D}, \tag{7}$$

or other measures that account for element D of the contingency table (e.g., Doswell et al. 1990). As always, it is essential to recall that there is no universal approach to verification but rather that the procedure selected needs to match the specific objectives of the study.

Regardless, some important verification concepts can be readily visualized in the diagram. Doswell (2004) summarized the duality of error, an idea elaborated by Hammond (1996) within the context of meteorological forecasts, by stating that ''at a given level of forecast accuracy for the system, false negatives can only be reduced by increasing the false positives and vice versa.'' Consider a fixed level of accuracy, as measured by the CSI, of, say, 0.30. For bias = 1.0, POD and SR are equal. Moving along the 0.30 contour in Fig. 2, it is clear that as SR increases, POD decreases and vice versa; that is, the slope of the contours is always negative. At the same time, of course, the bias increases or decreases in a completely determined fashion.

Second, it is important to understand the influence of sampling variability on the scores and, in order to assess skill, to reference the forecasts to some baseline. For large samples, one can estimate the distribution of the probabilities about the true values of each entry in the contingency table based on the central limit theorem and Gaussian statistics. In particular, the 95th percentile range for the probability of a given entry in the table goes as $N^{-0.5}$, where N is the sample size. It is possible to estimate variability in the scores owing to sampling without resorting to normal statistics using a form of

TABLE 1. Forecast quality measures for verification datasets. Additional measures (PCR, TSS, GS, and HSS) are provided for comparison (see text for details). The 95th percentile sampling error estimates for each of the measures are provided.

Dataset	POD	SR	Bias	CSI	PCR	TSS	GS	HSS
Network light snow	0.63 ± 0.07	0.69 ± 0.07	0.91	0.50 ± 0.05	0.77 ± 0.05	0.40 ± 0.12	0.26 ± 0.06	0.41 ± 0.12
Network heavy snow	0.61 ± 0.15	0.41 ± 0.10	1.49	0.32 ± 0.07	0.90 ± 0.02	0.51 ± 0.17	0.27 ± 0.07	0.42 ± 0.14
48-h convective occurrence	0.82 ± 0.08	0.94 ± 0.10	0.87	0.78 ± 0.08	0.94 ± 0.10	0.75 ± 0.18	0.60 ± 0.09	0.75 ± 0.18
SPC severe	0.49	0.69	0.71	0.40	NA	NA	NA	NA
SPC tornado	0.48	0.54	0.89	0.34	NA	NA	NA	NA
NWS TAF	0.65 ± 0.00	0.60 ± 0.00	1.08	0.46 ± 0.00	0.96 ± 0.00	0.61 ± 0.00	0.42 ± 0.00	0.59 ± 0.00
MOS TAF	0.54 ± 0.00	0.54 ± 0.00	1.01	0.37 ± 0.00	0.96 ± 0.00	0.50 ± 0.00	0.33 ± 0.00	0.50 ± 0.00
HPC precipitation (warm)	0.24	0.30	0.78	0.15	NA	NA	NA	NA
HPC precipitation (cold)	0.37	0.34	1.10	0.21	NA	NA	NA	NA
Reforecast positive 5 day	0.24 ± 0.01	0.28 ± 0.01	0.85	0.15 ± 0.01	0.82 ± 0.00	0.06 ± 0.01	0.03 ± 0.00	0.06 ± 0.01
Reforecast positive 7 day	0.21 ± 0.01	0.27 ± 0.01	0.78	0.14 ± 0.00	0.84 ± 0.01	0.05 ± 0.01	0.03 ± 0.00	0.05 ± 0.01
Reforecast positive 10 day	0.14 ± 0.01	0.29 ± 0.01	0.47	0.10 ± 0.01	0.90 ± 0.01	0.04 ± 0.01	0.03 ± 0.00	0.05 ± 0.01
Reforecast negative 5 day	0.30 ± 0.01	0.26 ± 0.01	1.15	0.16 ± 0.01	0.76 ± 0.00	0.06 ± 0.01	0.03 ± 0.00	0.06 ± 0.01
Reforecast negative 7 day	0.27 ± 0.01	0.26 ± 0.01	1.05	0.15 ± 0.01	0.78 ± 0.00	0.06 ± 0.01	0.03 ± 0.00	0.05 ± 0.01
Reforecast negative 10 day	0.21 ± 0.01	0.26 ± 0.01	0.80	0.13 ± 0.01	0.83 ± 0.01	0.04 ± 0.01	0.02 ± 0.00	0.05 ± 0.01

bootstrapping. The process is simple: create a new sample of the same size as the original with the same number of observed ($A + C$) and forecast ($A + B$) ''yes'' entries, and the same number of observed ($B + D$) and forecast ($C + D$) ''no'' entries. However, the sequence is otherwise random and in this sense represents a kind of climatological forecasting. By computing the accuracy measures for 1000 such samples, and sorting and extracting the 25th and 975th values, one can construct the 95th percentile range for the measure of interest. The sampling variability so defined is represented in Fig. 2 for POD and SR by the ''crosshairs'' and this can be used to see the effect on CSI (note: the severe convection and heavy rainfall forecasts, for which we do not have the element D, are not represented in this way). Additionally, this procedure will provide a baseline measure of accuracy based on the sample frequency, which can then be compared against the forecasts of interest to provide a measure of skill (also plotted in Fig. 2).

For the short-range forecast data (Fig. 2), most of the points are clustered toward the center of the diagram, relatively close to the bias = 1.0 line. Moving along this line from the lower left to the upper right of the diagram indicates an increase in absolute forecast accuracy. As discussed in Hammond (1996) and Doswell (2004), the aim of research and forecaster training is to increase the accuracy, that is, to move toward the upper right of the diagram. Optimal increases in accuracy are obtained by moving at 45°, that is, by maintaining unbiased forecasts through simultaneous increases in detection and reductions in false positives.

Next, we turn to several illustrative examples (see Table 1 for the scores and the appendix for the contingency table elements). Roebber et al. (2003) develop a neural network procedure by which, given the vertical profiles of temperature and moisture, surface wind, and liquid equivalent precipitation amount, one can diagnose snowfall density. Such information is crucial to making accurate forecasts of snow accumulation. The value of this procedure was studied within the context of municipal snow removal for the 2004–05 and 2005–06 cold seasons (Roebber et al. 2007) and in winter 2007–08 was adopted by the National Oceanic and Atmospheric Administration's Hydrometeorological Prediction Center (NOAA/HPC). Considerable skill for this task is indicated by the separation of the neural network and sample frequency scores for heavy and light snow densities (Fig. 2; Tables 1 and A1). Despite similar PODs, however, the neural network has higher accuracy for light snow density owing to the smaller bias, a propitious result since snowfall accumulation will be larger for lighter snow densities for a given amount of liquid equivalent. It is also evident that heavy snow density in the test sample is a relatively rare event, composing only 10% of the total cases, compared to 45% for light snow density.

Fowle and Roebber (2003) used a real-time mesoscale model with 6-km grid spacing to produce explicit forecasts of convective occurrence and mode in southern Wisconsin. These forecasts were then verified using radar reflectivity data. The surprisingly high accuracy for 48-h forecasts of convective initiation in the target region is evident (Fig. 2; Tables 1 and A2). The data show that the very few false alarms occur at the cost of a slight underforecasting of actual convective events, which composed 44% of the cases. In this instance, the accuracy of the model in anticipating nonevents is important as well, showing a 94% correct rejection.

Next, we consider NOAA/HPC heavy precipitation forecasts, defined as liquid equivalent precipitation in

excess of 12.5 mm within a 6-h period at a site. Raw forecast and observed precipitation data were obtained from the NOAA/HPC National Precipitation Verification Unit Quantitative Precipitation Forecast (NPVU-QPF) verification dataset for January 2001–July 2005 for all forecasts within 1000 km of HPC. These data contain 32-km gridded observations and HPC forecasts at time ranges of 6–24 h (see the Web site http://www.hpc.ncep.noaa.gov/npvu/help/remap.shtml for details). All 6-h observed or forecast precipitation events in excess of 12.5 mm are included in the analysis. A stratification based on cold season only forecasts (October–April), when forecast accuracy tends to be somewhat higher (see www.hpc.ncep.noaa.gov) is included.

The data (Fig. 2; Tables 1 and A3) demonstrate the meteorological challenge of forecasting heavy precipitation, a problem that is exacerbated in the warm season when convection dominates. It is notable that much of the accuracy increase from the warm to cold season is attributable to increased POD, which is again a signature of the change from thunderstorms to precipitation systems that organize on larger scales.

Significant severe weather reports for 2000–04 from the NOAA/Storm Prediction Center (SPC) Severe Thunderstorm Database are also examined. These reports are subdivided into nontornadic and tornadic events. Following Hales (1988), a nontornadic (hail or wind) event is considered significant if it meets any of the following criteria: 33 m s^{-1} or greater wind, 5-cm diameter or greater hail, one or more fatalities, three or more injuries, or damages in excess of $50,000. A tornadic event is considered significant if it meets any of these criteria: F1 or greater tornado, one or more fatalities, three or more injuries, or damages in excess of $50,000 [the tornadic criteria are relaxed to include F1 tornadoes, since these observations are still considered to be relatively robust; Verbout et al. (2006)].

Forecast watches (type, valid period, and latitude–longitude of the bounding rectangle) are obtained from a database maintained by the SPC. A severe watch is verified if any significant nontornadic severe weather report occurs within its boundaries during the valid time, while tornado watches are verified with reports of significant tornadoes within its boundaries during the specified interval. All cases within 1000 km of SPC are examined. Note that this verification procedure means that an event at a given date and time inside a watch box counts as a ''hit'' but that N events outside of a box will count as N ''misses.''

There is a slight increase in accuracy for SPC forecasts of significant severe weather relative to tornadic forecasts (Fig. 2; Tables 1 and A4). Event detection is the same, however, so this difference is entirely owing to a reduction in false alarms. Despite this, there is a slight underforecasting bias for both types of forecasts. The relative abundance of false positives for tornadic storms may represent the perceived penalty of missing a weather event, with the risk of casualties noted by Doswell (2004). An alternative explanation, which we cannot discount, however, is that the underforecasting bias is an artifact of the above verification method that can result in multiple misses for a given box.

Terminal aviation forecasts (TAFs) issued by all available NWS WFOs in the CONUS for 2003–05 are examined. Scheduled-only TAFs were verified using the ''operational impact scheme for flight category'' in the first 6 h of the TAF. A 2×2 contingency analysis for critical threshold data (ceiling <1000 ft or visibility <3 mi versus ceilings ≥ 1000 ft and visibility ≥ 3 mi) is used. As a reference, these forecasts are compared to those derived from the model output statistics (MOS). Notable is the lack of bias in both sets of TAF forecasts (Fig. 2; Tables 1 and A5), but the substantial increase in accuracy of the NWS forecasts relative to MOS (the sampling variability is negligible for these large samples; see Table 1). Since the forecasts are relatively unbiased, the human forecast skill in the TAF forecasts is obtained both through increases in detection and reductions in false positives. Note that this again represents a rare-event type of forecast, with only 8% of the sample meeting the critical threshold criteria.

Illustration of the temporal effects on forecast quality can also be accomplished with the verification diagram. A natural application is medium-range dichotomous forecasts of a quantity of interest; a ''trajectory'' is traced out in the diagram space corresponding to the evolution of forecast quality, trending toward climatology at the greatest range. To demonstrate this aspect, the NOAA Climate Diagnostic Center's 15-member ensemble forecast archive (Hamill et al. 2006), generated from the NOAA/NCEP Global Spectral Model (GSM; specifically, the operational version of the MRF that was in place from January to June 1998) is considered. All 500-hPa forecasts for December–February for 1979–2003 at 5-, 7-, and 10-day ranges are examined in the hemispheric band from 35° to 45°N. At every 10° longitude, a 500-hPa anomaly is computed relative to the 1979–2003 mean. If the ensemble forecast or observed anomaly is at least one standard deviation from the climate normal, then it is characterized as a positive or negative anomaly depending on the sign of the departure (Fig. 3; Tables 1 and A6). Based on these criteria, slightly more than half of the verified dates and longitudes exhibit no anomalies and for a given positive or negative forecast, only about 22% of the sample would be considered an event.

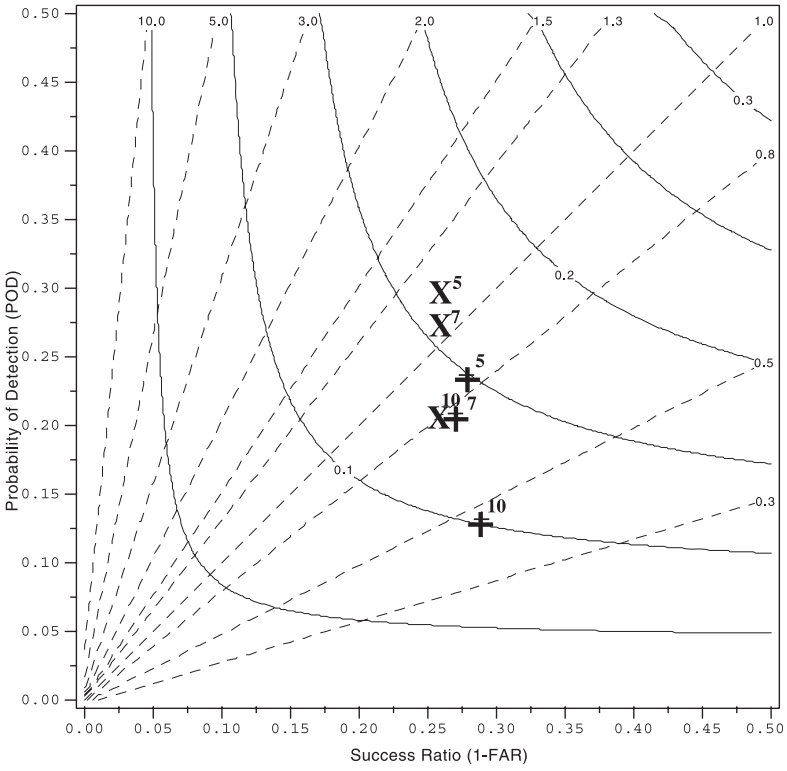

FIG. 3. Performance diagram as in Fig. 2, but for the ensemble forecast positive (plus symbol) and negative (X) 500-hPa height anomalies. Shown are three forecast ranges, as indicated by the superscripts: 5, 7, and 10 days. Sampling variability, which is small, is omitted in this diagram. See text for details.

Two aspects of the verification are immediately obvious. First, all of the forecasts exhibit relatively low skill (note that the climatological forecast, by definition, is no anomaly and therefore has a POD and a CSI of zero) but the negative anomaly forecasts are superior, owing to a higher POD. Second, the loss of skill with increased forecast range in both cases is largely owing to a decreased POD, while SR remains relatively constant from 0.29 to 0.26 and percent correct rejections increase (note that the loss of detection and skill culminates in a climatological forecast of zero skill and 100% correct rejections for anomalies of either sign). These differential performance issues between positive and negative anomalies are not highlighted by metrics like the TSS, GS, and HSS (Table 1) and may suggest a clue toward GSM physics that bears further attention.

3. Summary

Measures-oriented verification methods remain in considerable use in many areas of forecasting, most particularly in the evaluation of dichotomous (yes–no) forecasts. In this note, we show that POD, FAR, bias, and CSI are geometrically related and consequently can be plotted in a single diagram. This visual representation is in general preferable to simple tables for ease of interpreting the statistics. Examples showing verification results plotted in this manner are provided for a range of forecasting applications, including snowfall, convective initiation within an area, severe convective type (nontornadic and tornadic), heavy rainfall, terminal aviation forecasts (ceiling and visibility), and medium-range ensemble 500-hPa height anomalies. While such examinations cannot reveal the origins of skill, they can point the way toward the more detailed investigations that are required to advance forecasting in the face of ongoing scientific, technical, and political obstacles.

Acknowledgments. We are indebted to NOAA/HPC and SPC for making some of these data available for analysis. TAF data were provided by the NWSEO. The author thanks Melissa (Butt) Schumann for her assistance in data analysis. Three anonymous reviewers provided some helpful suggestions on an earlier version of this manuscript.

APPENDIX

Forecast Contingency Tables

TABLE A1. Roebber et al. (2003) neural network snow density diagnoses for light and heavy snow.

		Observed light snow	
		Yes	No
Network light snow	Yes	95	42
	No	55	141
		Observed heavy snow	
		Yes	No
Network heavy snow	Yes	20	29
	No	13	271

TABLE A2. Real-time mesoscale model 48-h forecasts of convective occurrence as reported in Fowle and Roebber (2003).

		Observed convection	
		Yes	No
Model forecast convection	Yes	62	4
	No	14	61

TABLE A3. NOAA/HPC 6–24-h heavy precipitation forecasts for warm and cold seasons (see text for details).

		Observed precipitation (warm)	
		Yes	No
HPC forecast precipitation (warm)	Yes	18 282	42 405
	No	59 652	NA
		Observed precipitation (cold)	
		Yes	No
HPC forecast precipitation (cold)	Yes	11 934	23 538
	No	20 299	NA

TABLE A4. NOAA/SPC significant severe weather and tornadic forecasts for 2000–04 (see text for details).

		Observed severe convection	
		Yes	No
SPC forecast severe convection	Yes	4588	2039
	No	4811	NA
		Observed significant tornadoes	
		Yes	No
SPC forecast significant tornadoes	Yes	679	572
	No	735	NA

TABLE A5. NWS and MOS TAFs verified for critical threshold data (ceiling and visibility) for 2003–05 (see text for details).

		Observed critical threshold	
		Yes	No
NWS forecast critical threshold	Yes	805 863	529 003
	No	432 651	1 326 1243
		Observed critical threshold	
		Yes	No
MOS forecast critical threshold	Yes	673 324	580 223
	No	565 191	13 210 023

TABLE A6. Reforecast 15-member ensemble forecasts for DJF 1979–2003 of positive and negative 500-hPa anomalies in the latitude band of 35°–45°N. Forecast ranges are 5, 7, and 10 days (see text for details).

		Observed 500-hPa positive	
		Yes	No
5 day forecast 500-hPa positive	Yes	4463	11 457
	No	14 234	52 610
		Observed 500-hPa negative	
		Yes	No
5 day forecast 500-hPa negative	Yes	5452	15 410
	No	12 761	49 141
		Observed 500-hPa positive	
		Yes	No
7 day forecast 500-hPa positive	Yes	3950	10 561
	No	14 747	53 506
		Observed 500-hPa negative	
		Yes	No
7 day forecast 500-hPa negative	Yes	4999	14 183
	No	13 214	50 368
		Observed 500-hPa positive	
		Yes	No
10 day forecast 500-hPa positive	Yes	2527	6153
	No	16 137	57 838
		Observed 500-hPa negative	
		Yes	No
10 day forecast 500-hPa negative	Yes	3805	10 750
	No	14 391	53 709

REFERENCES

Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting,* **11,** 288–303.

Doswell, C. A., III, 2004: Weather forecasting by humans—Heuristics and decision making. *Wea. Forecasting,* **19,** 1115–1126.

——, 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.

Fowle, M. A., and P. J. Roebber, 2003: Short-range (0–48 h) numerical prediction of convective occurrence, mode, and location. *Wea. Forecasting,* **18,** 782–794.

Gribbin, J. R., and M. Gribbin, 2004: *FitzRoy: The Remarkable Story of Darwin's Captain and the Invention of the Weather Forecast*. Yale University Press, 336 pp.

Hales, J. E., 1988: Improving the watch/warning system through the use of significant event data. Preprints, *15th Conf. on Severe Local Storms*, Baltimore, MD, Amer. Meteor. Soc., 165–168.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Hammond, K. R., 1996: *Human Judgment and Social Policy*. Oxford University Press, 436 pp.

Jolliffe, I.T., D.B. Stephenson, Eds., 2003: *Forecast Verification. A Practitione''s Guide in Atmospheric Science*. John Wiley and Sons, 240 pp.

Katz, R.W., A.H. Murphy, Eds., 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp.

Lambert, S. J., and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dyn.,* **17,** 83–106.

Morss, R. E., J. K. Lazo, B. G. Brown, H. E. Brooks, P. T. Ganderton, and B. N. Mills, 2008: Societal and economic research and applications for weather forecasts: Priorities for the North American THORPEX program. *Bull. Amer. Meteor. Soc.,* **89,** 335–346.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

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

——, B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4,** 485–501.

Ramage, C. S., 1993: Forecasting in meteorology. *Bull. Amer. Meteor. Soc.,* **74,** 1863–1871.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting,* **18,** 264–287.

——, M. R. Butt, S. J. Reinke, and T. J. Grafenauer, 2007: Real-time forecasting of snowfall using a neural network. *Wea. Forecasting,* **22,** 676–684.

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

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2004. *Wea. Forecasting,* **21,** 86–93.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Willis, E. P., and W. H. Hooke, 2006: Cleveland Abbe and American meteorology, 1871–1901. *Bull. Amer. Meteor. Soc.,* **87,** 315–326.