FORECASTER'S FORUM

Discussion of Verification Concepts in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*

BOB GLAHN

Meteorological Development Laboratory, National Weather Service, Silver Spring, Maryland

16 December 2003 and 9 February 2004

The book *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, edited by Jolliffe and Stephenson (2003, hereafter JS03), fills a void in verification of meteorological and climate forecasts. While a number of books on aspects of statistics related to meteorology and climatology (e.g., Wilks 1995) discuss verification, this complete book is devoted to the subject. The book comprises a fairly tightly coupled set of chapters written by generally well-known experts, in some cases perhaps more so in Europe than North America, in verification and especially in the subjects of their particular chapters. In a book in which sections or chapters are written by different authors, one asks the following questions: 1) how well do the individual chapters read and present the material logically, accurately, and comprehensively; and 2) how well do the chapters relate to one another and address the full subject of the book? Regarding the first question, JS03 gets high marks for most chapters. On the second question, JS03 is better than many, although the editors have not suppressed individuality enough in some instances for it to read like a fully cohesive book. JS03 is a voluminously referenced and well-indexed survey of what is known about, and a historical account of, verification and the related topic evaluation as it exists in the meteorological literature. The editors have put much emphasis on standardizing mathematical notation throughout, and were quite successful—an achievement in itself. While the methods presented can be applicable to most any forecasting problem, the discussion and examples are tied to weather and climate forecasting as acknowledged by JS03 (preface), which hardly translates into the full scope of ''atmospheric science.''

I have been interested in and involved with verification even before my entry into the U.S. Weather Bureau in 1958. In the Alaskan Weather Center of the U.S. Air Force, as the first numerical weather prediction (NWP) ''progs'' were rolling out, we were using a score similar to the S1 score (JS03, p. 129; Teweles and Wobus 1954). Roger Allen (for whom I worked for several years) and Jack Thompson hired me, and my office was just down the hall from Glenn Brier and Thompson. All three had recently published what has turned out to be landmark papers (Brier 1950; Thompson 1952; Thompson and Brier 1955; Brier and Allen 1951); all except Thompson and Brier are referenced in JS03. Over the years, I have watched the verification literature grow to what it is today. I certainly agree with JS03 that ''Allan Murphy had a major impact on the theory and practice of forecast verification'' (p. 3). Murph was a prolific writer, maintaining over long periods a paper a month. He collaborated with many others of renown and touched on most subjects relating to forecast verification. The one topic with which he had not gotten entirely comfortable was forecasts of spatial fields (Allan Murphy, personal communication), although he and Ed Epstein defined a skill score for model verification (Murphy and Epstein 1989). Perhaps the single most important paper he coauthored was the landmark paper, ''A general framework for forecast verification'' (Murphy and Winkler 1987), mentioned by JS03 in chapter 1.

A possible runner-up in importance to the Murphy and Winkler paper in the meteorological verification literature was the introduction of the relative operating characteristic (ROC) into meteorology. While Murph embraced this concept, it was first brought into the meteorological literature by Ian Mason (1980, 1982a,b), who reported and built upon the work of John Swets (1973). John and Ian were two of the invitees to a workshop on probabilistic weather forecasts in 1983 at Timberline Lodge on Mount Hood, Oregon, organized by Murphy. ROC has not played as major a role in the past as such scores as versions of the skill score and threat

Corresponding author address: Dr. Harry R. Glahn, Meteorological Department Laboratory, National Weather Service, W/OST2, 1325 East–West Highway, Silver Spring, MD 20910.
E-mail: harry.glahn@noaa.gov

score (sometimes with different names), but it is beginning to come to the forefront with the recognition and use of probability information. Even the terminology base rate, hit rate, and false alarm rate have come into prominence in meteorological forecast verification largely through the influence of ROC. For instance, JS03 in the definition for hit rate states that it is ''Also known as probability of detection in older literature.'' I would counter that most readers and developers associated with atmospheric science are more familiar with ''probability of detection'' than they are with ''hit rate.'' Maybe that is because they, too, are ''older.''

I have identified a number of recurring themes or central ideas in JS03 mentioned below:

Finley's tornado forecasts. The now-famous 2 × 2 table of Finley's (1884) yes/no tornado forecasts is introduced on page 1 and is discussed several times. The table even appears almost subliminally on the cover. JS03 states ''. . . there is a surprisingly large number of ways in which the numbers in the four cells . . . can be combined to give measures of the quality of the forecasts.''

Verification presented from a developer's viewpoint. Much of the discussion seems to have as an objective developing or improving a forecast system rather than judging the, possibly comparative, goodness of a set of forecasts. While both aspects are important, JS03 does not clearly make the distinction, and I would have expected concentration to be heavily on the latter rather than the former.

Strong emphasis on the ROC and its associated terminology hit rate, H, *and false alarm rate,* F. While other ways to evaluate forecasts [e.g., computation of scores, such as mean absolute error (MAE)] are treated throughout the book, the ROC gets a very strong play. Albeit an important concept, it has a major deficiency—it does not consider calibration, and poorly calibrated forecasts may be judged to be as good as well-calibrated forecasts. This is stated in JS03 in some contexts, but is not emphasized, and when it is mentioned, it is usually dismissed with the suggestion to recalibrate, in keeping with the development theme.

Probability forecasts. In agreement with the recent American Meteorological Society (AMS) statement on probability forecasts (AMS 2002), JS03 recommends the use of probability forecasts and emphasizes their potential value to customers over nonprobabilistic forecasts.

Ensembles. The examples are, beside Finley's forecasts, in connection with climate or ensembles. Climate forecasts can provide good data in many contexts, but ensembles are overplayed almost to the nonrecognition that there are other ways to make probability forecasts.

Contributions of Allan Murphy. While it is no surprise to those interested in verification that Murphy's work would get many citations, the number and diversity is so great that it is a dominant thread in JS03.

Attribution to other authors. JS03 provides many citations to previous works, which can be very helpful to those delving into details of verification and evaluation of meteorological and climate forecasts.

In chapter 1, the editors reiterate Brier and Allen's (1951) reasons for verification; use their terms ''economic,'' ''administrative,'' and ''scientific,'' and note that a common theme is that any verification scheme needs to be informative (p. 4). They note that it is highly desirable that the verification system be objective; they examine various scores according to attributes reliability, resolution, discrimination, and sharpness, as suggested by Murphy and Winkler (1987), and for ''goodness'' of which Murphy (1993) identified three types—consistency, quality (e.g., accuracy or skill), and value (utility); and they note that in order to quantify the value of a forecast, a baseline is needed, and that persistence, climatology,[1] and randomization are common baselines. I might add that the well-established objective method of Model Output Statistics (MOS) produces an important and more competitive baseline for many forecasts, especially in the National Weather Service in the United States; but, this is not mentioned in JS03.

While the idea of a baseline is important and seemingly a simple concept, even climatic forecasts as a baseline need more definition, because different ''definitions'' can give quite different results. For instance, in verifying temperature forecasts over a season, the mean temperature (climatic mean) over the season would be a poor baseline. One should rather use monthly means or some simple low-frequency curve fit to the data over the same seasonal extent. Even so, the question of using the sample frequencies of categories versus longer-term relative frequencies usually is not a given. For instance, Bob Livezey (in JS03, p. 78) states ''. . . the exclusive use of sample probabilities (observed frequencies) of categories of the forecast/observation set being verified is recommended, rather than the use of historical data. The only exception to this is for the case where statistics are stationary and very well estimated.'' However, as with many verifications, the purpose comes into play. If one is comparing a set of subjective temperature forecasts with the baseline available to the forecaster when the forecasts are being made, the baseline is the historical record, not the mean of the time series yet to be observed, regardless of the stationarity of the time series. (Extreme nonstationarity would indicate that climatic forecasts were inappropriate as a baseline, but this is usually not known when the forecasts are being made,

[1] Some would say that the term climatology is not appropriate here according to its strict definition, and that climatic forecasts, climatological forecasts, or something similar, would be more appropriate [see the *AMS Glossary of Meteorology* (Glickman 2000)].

so climatic forecasts is the available baseline that is used.) In any case, the usual ''skill'' scores computed on multidimensional tables do generally base skill on the sample.

It is also worth noting that the reduction of variance used in regression and predictor selection is relative to the overall mean of the sample. Therefore, a very high ''score'' can be obtained by getting only the seasonal variance right. I note that climatological forecasts, as used by Murphy and Epstein (1989), refer to long term and not sample, a departure from Bob Livezey's recommendation concerning categorical forecasts and implied by Deque in JS03's chapter 5.

While development of objective forecasting systems is not the subject of the book and gets little direct treatment (other than ensembles), the authors do note that artificial skill is a danger in developing a forecasting system and emphasize cross validation and separate training and test datasets. In this regard, I note the changing terminology from what was used when objective techniques were assuming prominence (e.g., Thompson 1950; Allen and Vernon 1951). ''Dependent'' or ''development'' data were used to develop systems, rather than ''training'' data, and the data to judge whether the system would hold up on new data were called ''test'' or ''independent'' data. I perceive that the term training has been highly influenced, unfortunately in my view, by the relatively modest development of systems requiring iterative solutions (e.g., neural networks; Marzban and Stumpf 1996), brought into the U.S. meteorological literature under the name of adaptive logic by Hu and Root (1964), rather than more analytic solutions (e.g., regression). The terminology and digression from verification per se is likely occasioned by some of the authors' backgrounds as developers of objective systems.

It is curious that Potts (in JS03, p. 13) calls the variable for which the forecasts are formulated the ''predictand'' and seems to justify that terminology by implying that all forecasts are made by ''forecasting systems.'' Strictly speaking, that may be correct, but the preponderance of forecasts, other than those made by NWP, are made subjectively by forecasters, and the variable that they are forecasting is generally not thought of as a predictand. This term comes from statistical objective systems, dating back to 1949 or before.[2] Even in ''objective'' NWP systems, the variables being forecast are not, to my knowledge, thought of as predictands. In the *AMS Glossary of Meteorology* (Glickman 2000), predictand is defined only in terms of regression (p. 594, 641).

Potts states that a predictand (again, her use of the term) can be either deterministic or probabilistic. This is now a common use of the term ''deterministic,'' but it carries more baggage than it is worth in my estimation. The term just means ''nonprobabilistic'' and not that there is necessarily any fundamental law that would ''determine'' a specific and correct value. But it has wide acceptance, probably for want of another more appropriate term, and will likely not fade (that is my probabilistic forecast). I like the term ''definite'' used by Drosdowsky and Zhang (in JS03, p. 121) or ''definitive'' better than deterministic.

Potts states (p. 14), ''A deterministic forecast is really just a special case of a probabilistic forecast in which a probability of unity is assigned to one of the categories and zero to the others.'' However, the editors state in chapter 9 (p. 192), ''. . . deterministic point forecasts are not perfectly sharp forecasts with probabilities equal to 1 and 0, but instead they should be assigned unknown sharpness.'' Even though these statements are brought together (only) in the glossary, it is not clear what view the book espouses.

Potts (p. 22) states that the sample variance is an unbiased estimate of the population variance, and defines the sample variance as

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1)$$

with $n - 1$ in the denominator. This also has become commonplace both in meteorology (e.g., Wilks 1995, p. 25) and statistics (e.g., Neter and Wasserman 1974, p. 10) texts, but I prefer the definition of sample variance having ''n'' in the denominator,

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad (2)$$

as was earlier done (Panofsky and Brier 1958, p. 26; Kendall and Stuart 1961, p. 4; Mood 1950, p. 132; Mode 1951, p. 64; Brooks and Carruthers 1953, p. 40; Klugh 1974, p. 53; Johnson and Jackson 1959, p. 32; and Underwood, et al. 1954, p. 67), with (1) being the unbiased estimate of the population parameter. The change in definition was already taking place in the 1950s (Panofsky and Brier, op. cit., footnote p. 26); the reason is unclear to me. Why would the ''mean'' of the squares about the sample mean be found by dividing by $n - 1$? There are, as Kendall and Stuart (1961, op. cit., p. 4) state, ''. . . reasons for preferring (1) to (2) as an estimator of the parent variance, notwithstanding the fact that the latter *is* (italics mine) the sample variance.'' What if the sample consisted of the entire population; why would (1) be used? One must keep straight whether a *sample statistic* is being *calculated* or a *population*

[2] Lorenz (1956), in his landmark paper on empirical orthogonal functions, uses the terms predictand and predictor. These terms were not in vogue in the 1940s when largely graphical methods of objective forecasting systems were developed and documented in unpublished U.S. Weather Bureau Research Papers (available from the National Oceanic and Atmospheric Administrations's library), and published in the *Monthly Weather Review*. Even Bob White, one of the first to apply regression to weather forecasting, did not use the term in early papers (e.g., White and Galligan 1956). However, Gringorten (1949), as early as 1949 in a statistical forecasting study using the sorting of punched cards, carefully defines and uses both terms.

parameter is being *estimated*. I believe this definition of (1) leads to an inconsistency in JS03's Eqs. (2.2) and (5.14), both of which purport to be the definition of skewness. If one is a sample and one an estimate of the population, it is not clearly stated. Also, Eq. (5.14) seems to not agree with some other texts [e.g., Wilks 1995, Eq. (3.8)].

Ian Mason goes into great detail in chapter 3 dealing with the two-category event, and continues a theme of the book in discussing this situation in terms of Finley's (1884) tornado forecasts. Up until Ian introduced the concept of the ROC into the meteorological literature, the verification of binary events centered around scores (e.g., Heidke skill, Hanssen and Kuipers, critical success index) computed on the 2×2 contingency table. If computing these scores can be considered a ''method,'' then one can agree with Ian, ''It is probably fair to say that there was very little change in verification practice for deterministic binary forecasts until the 1980's, with the introduction of methods from signal detection theory (SDT) . . . and the development of a general framework for forecast verification by Murphy and Winkler (1987).''

As mentioned earlier, in some respects, the ROC methodology is, at present, a close runner-up to the Murphy and Winkler (1987) paper in influence, and that influence will likely be felt more with time, at least until its major deficiency of not considering reliability is fully appreciated. In this book, the terminology brought from the ROC methodology ''base rate,'' ''hit rate,'' and ''false alarm rate'' is predominantly used, and most scores are in, or put into, those terms. If these terms prevail, it will be because of the ROC influence. We can be thankful the terms prefigurance and postagreement (Panofsky and Brier 1958) are not mentioned.

I find it curious that there are several places (p. 50, 53, 55, 70, 73) that negative skill (a set of binary forecasts that do worse than the baseline) seems to be no problem to the authors—just reverse the labels, and the negative skill becomes positive. Well, so! But usually we do not have the luxury of changing the forecasts we are verifying. Evidently, this statement is from the perspective, as mentioned earlier, of developing an objective system and the influence of ROC; but, the book is about verification, which involves determining the correspondence of the forecasts and the ''observations,'' not about switching labels at the end. Belaboring this point is not useful to a ''practitioner'' of verification.

ROC is given good treatment, covering its relationship to type-1 and -2 errors in hypothesis testing. The parametric (or modeled) area under the ROC curve is described along with the associated discrimination distance. One should keep in mind that this model seems to work in many instances without strong analytic justification. Not explicitly noted is the fact the ROC measures or graphs only ''discrimination'' ability of the set of forecasts, and does not rely on the forecasts being well calibrated (i.e., reliable). The ROC methodology was developed to discriminate signal from noise, and a ''signal'' or ''forecast'' not being calibrated may not be a terrible disadvantage in some applications, and even, perhaps, in developing an objective system. However, I believe this measure of discrimination (only) is currently being overemphasized when one is concerned with a full measure of the correspondence between forecasts and observations. While past forecasts can be calibrated, and this calibration may hold on similar future forecasts, ''mislabeled'' forecasts provided to a user, as discussed in JS03's chapter 7, would likely not be useful to her.

JS03 states, ''The resolution component of the Brier score and the (area under the) ROC curve therefore often provide very similar information.'' They state, ''A potential advantage of skill measures such as the ROC area is that they are directly related to a decision-theoretic approach and so can be easily related to the economic value of probability forecasts for forecasts users.'' It is not clear to me how reliability (calibration), which is generally ignored by ROC, can not be crucial in determining the actual (rather than potential, if properly calibrated) economic value of forecasts. And, is the area under the curve a ''skill measure'' or an ''accuracy'' measure? Or neither? JS03, in chapter 8, calls the area under the curve a ''summary skill measure,'' but then goes on to formulate a skill score based on the ROC area. Ian Mason makes an interesting reference to Finley's (1884) tornado forecasts, and concludes, based on the ROC methodology, ''. . . at 95% level . . . Finley's forecasts did have some skill!''

Richardson also addresses the ROC in chapter 8 and, in contrast to Mason in chapter 3 where only the *modeled* area under the curve A_z is discussed, goes to some length to discuss the *actual* area A under the curve when points are plotted on the hit rate–false alarm rate axes. He also defines a ROC skill score as ROCSS $= 2A - 1$, which ranges from 0 for no skill to 1 for perfect forecasts. Nothing is said about the possibility that ROCSS could be negative, which may go along with Mason (p. 70), ''ROC points below the diagonal represent the same level of skillful performance as they would if reflected about the diagonal. If a forecasting system produces ROC points in this area, the forecasts are mislabeled.'' This is also hinted at by Richardson (p. 175), ''If the forecasts are not reliable, then the threshold should be adjusted . . . the calibration procedure . . . makes this adjustment . . .'' Again, quite so for developing and adjusting a forecast system for future use, but *not* for verifying or evaluating a set of existing forecasts.

The argument is sometimes made that if biased probability forecasts are given to a user, he/she will find that out and ''recalibrate'' or change his/her threshold. I would counter that, unless the ''system'' making the probability forecasts is objective and its characteristics can be expected to hold in the future, the user is playing a dangerous game. A provider of such forecasts ought

to also notice the bias and correct it, making the adjustment made by the user invalid. It is true, a user may set a threshold such that, for instance, more ''forecasts'' of severe weather are made than actually occur (a bias of categorical forecasts) because of his/her utility matrix, but there is no excuse for providing a user with biased *probability* forecasts.

In whatever chapter the ROC is described, both the actual (from plotted points) and the modeled area should be discussed (in that order), not sequestered for the reader to attempt to coalesce. Some authors prefer the ''modeled'' diagram and proclaim that connecting consecutive points by straight lines underestimates the area under the curve. Quite so, but if the system being verified produced, or is capable of producing, only the specific probabilities associated with the plotted points in the diagram, it is not really appropriate to connect them *at all* except as an eye assist; if the points are very close together, it matters little how they are connected, and such connection is reasonable.

Bob Livezey discusses multicategory events and reviews various scores, but soon mentions that most scores are deficient when compared to the Gandin and Murphy (1992) ''equitable'' family of scores. Although the Heidke and Peirce skill scores are both equitable, they have the undesirable properties of depending on the forecast distribution, and not utilizing off-diagonal elements in the contingency table. Bob Livezey also discusses the relatively new LEPSCAT score and sampling variability of the contingency table and skill scores, but a major thrust of the chapter leads to the family of Gandin and Murphy scores, how they can be constructed, and the Gerrity (1992) score (GS), one of the family, is recommended as the preferred one.

Gandin and Murphy scores are based on a reward (or penalty) value for each cell in the contingency table. This is the same concept that is used for determining the ''value'' of a set of forecasts, in contrast to skill, where the rewards and penalties are applied to a particular operation and are known or can be estimated [see Miller and Starr (1960, 82–85) for an early example of using weather forecasts in decision making under risk]. The trick here, as a general skill problem, is how to determine the reward (or penalty) matrix. Conditions for equitability can be defined, but they by themselves are insufficient to determine the matrix, so certain further conditions (or restraints) are made. Gerrity set constraints that, as it turned out, gives his scores a remarkable property. A Gerrity score computed on the full k-cell table is the same as the arithmetic mean of all the $k - 1$ two-category Gerrity scores formed by combining categories on either side of the partitions between consecutive categories. Livezey notes this remarkable property and states, ''Because of its convenience and built-in consistency, the family of GS is recommended here as equitable scores for forecasts of ordinal categorical events.'' This appears to be a good choice, but one must still remember that there is necessarily a certain arbitrariness to the values in the cells of the defining table.

Deque in chapter 5 uses the term ''variable'' instead of the statistical developers term ''predictand'' preferred by Potts (p. 13). Although ''quantity'' is also used for variables such as temperature and pressure, I prefer to reserve quantity for something quantitative, not a substitute for a random variable; however, this usage of quantity has now become common in meteorological literature.

JS03 gives some, but minimum, treatment to the related topics sampling error, artificial skill, and significance testing. These are very important topics and deserve more consideration. ''Prediction interval'' is contrasted to ''confidence interval'' (p. 105), but no definitive explanation of the difference is given; verification as a regression problem is mentioned in chapter 2, and the discussion of the Pearson product moment correlation coefficient is here (p. 106) and provides an excellent opportunity to demonstrate the difference.

One can agree with Deque's statement, ''. . . it is desirable that the overall distribution of forecasts is similar to that of the observations, irrespective of their case-to-case relationship with the observations'' (p. 113). However, the statement, ''Before the forecasts are delivered to unsuspecting users, it is important to rescale (inflate) them'' (p. 114) can be questioned. A definition of ''inflate'' is not given, but has come to mean in many instances that defined for regression estimates by Klein et al. (1959) (no attribution in JS03), and may or may not be desirable. The mean square error skill score (p. 104) for inflated unbiased forecasts will be negative if the (Pearson product moment) correlation coefficient between noninflated forecasts and observations is <0.5 (Glahn and Allen 1966). That is, in developing the regression equation, if the reduction of variance is <0.25, inflated forecasts will have a larger mean square error than the sample mean. An unsuspecting user, having been given inflated forecasts, might expect them to be skillful!

It is interesting that one of the most challenging verification problems, that of dealing with spatial fields, is given relatively short treatment. The different anomaly correlation coefficients in the literature and S1 score (Teweles and Wobus 1954) are defined, and principal component analysis is introduced as a method of reducing dimensionality. Spatial rainfall forecasts are singled out as being especially challenging to verify.

Introducing a ''spacial'' dimension adds a truly new dimension to the complexity of verification. The basic question of ''What is a good forecast?'' becomes more difficult and may depend more heavily on the forecast's purpose. For instance, if the ''pattern'' is right, but displaced, is that a good forecast? While discussed in reference to rainfall forecasts, the question is pertinent for most all fields of ''weather'' variables. Novel approaches relying on translating patterns to get a good match are only briefly introduced. The last paragraph mentions

the possibility that the field can be verified at different spatial scales. Although pertinent references are given, one is left with a certain incompleteness regarding spatial fields.

My biggest disappointment with JS03 is their rolling the verification of probability forecasts into a chapter shared by ensemble forecasting. JS03 states (p. 155), ''Ensemble forecasting is now one of the most commonly used methods for generating probability forecasts that can take account of uncertainty in initial and final conditions.'' While ensemble forecasting is in its ascendancy, and the statement is true in terms of the uncertainty of the initial conditions estimated by data assimilation, precious little overall work has been done operationally with ensembles in a postprocessing probabilistic sense, except for the occurrence of precipitation, which, being binary, lends itself well to direct relative frequency treatment. It is *not* the most commonly used method of producing probability forecasts; rather, statistical postprocessing of single-model runs have produced a plethora of probabilistic guidance forecasts for *many* weather elements for many years, including probability of precipitation, type of precipitation (freezing, frozen, or liquid, and showers, drizzle, or rain), wind, cloud amount, ceiling height, and visibility. No reference is given to U.S. or Canadian work in the short range (0–10 days), both of which have been primary centers of postprocessing activity for many years. All of this information, along with, more recently, ensemble information, is provided to forecasters as guidance in making the ''official'' forecasts. In addition, probability of precipitation forecasts have been produced as official forecasts by the U.S. National Weather Service since 1966 (Hughes 1980). It is a disservice to the unsuspecting reader to suggest that we really need only be concerned with probability forecasts produced directly from ensembles. Unfortunately, the ''equating'' of probability forecasts and ensembles is carried over into chapter 8, where in the statement, ''The contrasting effects of ensemble size on the ROC area and Brier skill scores are examined in Section 8.5 . . . ,'' there is no recognition that ensembles are not the only game in town. JS03's editors should have forced a more balanced view. I even ask, why is a chapter on ensemble forecasting, which is a method of making forecasts, not verifying them, put into a book on verification? Why not include a chapter on regression, or discriminant analysis, both methods of producing probabilistic forecasts?

JS03 states ''. . . the two most important attributes of probability forecasts (are) referred to as reliability and resolution'' (p. 138), not news to the reader at this point. Later, they say that resolution is the *most* important attribute of a forecast system (p. 142). While these two statements are not quite contradictory, editing could have provided a more clear picture of the authors' views and more clearly differentiated a set of probability forecasts from a system that *could* produce reliable forecasts by recalibrating. The idea is that forecasts do not (or that a forecast system does not) really have to be reliable, one just has to calibrate so that they are. This is emphasized later (p. 163) and contributes to the perception that the book is oriented for a developer of systems, not for one who is going to verify or evaluate an *actual*, *unchangeable* set of forecasts. Both purposes of verification are important, but JS03 never clearly makes the distinction.

Chapter 8, written by Richardson, discusses the third type of goodness previously identified by Murphy, value or utility. Other measures associated with the correspondence of forecasts and observations (e.g., skill and accuracy) are not *directly* measures of the usefulness of forecasts to a user, although they are certainly related. This topic was brought into modern U.S. meteorological literature by Jack Thompson (1950, 1952; Thompson and Brier 1955). Thompson formulated the now-familiar cost–loss decision model,[3] and the same model has been described and discussed many times since, as Richardson indicates, and is summarized in this chapter. In keeping with a theme in JS03, hit rate and false alarm rate are brought into play, and the usefulness of the Peirce skill score and the Clayton skill score in this context are discussed. Through analysis, Richardson concludes, in concert with previous authors, ''. . . no single threshold probability (e.g., a deterministic forecast) will be optimal for a range of users with different cost/loss ratios—this is a strong motivation for providing probability rather than deterministic forecasts.''

If there is a single user with an identified cost of protection for ''adverse'' weather and loss if protection is not taken, then the utility for a set of forecasts for that user can be calculated. Also, if the *distribution* of users is known, the overall utility can be calculated *and* related to the Brier score, which Richardson shows. For instance, if the distribution of cost–loss ratios for users is uniform over the 0–1 range, then (per unit loss) ''the Brier skill score is the overall value'' (p. 183).

I found JS03 to be very interesting reading and will be useful, provided some of its concepts and statements are carefully evaluated for the specific use to be made of them. Perhaps this paper and likely follow-on dialogue will help improve the second edition.

REFERENCES

Allen, R. A., and E. M. Vernon, 1951: Objective weather forecasting. *Compendium of Meteorology*, T. F. Malone, Ed., Amer. Meteor. Soc., 796–813.

AMS, 2002: Enhancing weather information with probability forecasts. *Bull. Amer. Meteor. Soc.*, **83,** 450–452.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

——, and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology*, T. F. Malone, Ed., Amer. Meteor. Soc., 841–855.

[3] Anders Angstrom and others had discussed this same model in papers not well known until relatively recently (Liljas and Murphy 1994).

Brooks, C. E. P., and N. Carruthers, 1953: *Handbook of Statistical Methods in Meteorology*. Her Majesty's Stationery Office, 412 pp.
Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.,* **1,** 85–88.
Gandin, L. S., and A. H. Murphy, 1992: Equitable scores for categorical forecasts. *Mon. Wea. Rev.,* **120,** 361–370.
Gerrity, J. P., Jr., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.,* **120,** 2707–2712.
Glahn, H. R., and R. A. Allen, 1966: A note concerning the ''inflation'' of regression forecasts. *J. Appl. Meteor.,* **5,** 124–126.
Glickman, T., Ed., 2000: *Glossary of Meteorology*. 2d ed. Amer. Meteor. Soc., 755 pp.
Gringorten, I. I., 1949: A study in objective forecasting. *Bull. Amer. Meteor. Soc.,* **30,** 10–15.
Hu, M. J. C., and H. E. Root, 1964: An adaptive data processing system for weather forecasting. *J. Appl. Meteor.,* **3,** 513–523.
Hughes, L. A., 1980: Probability forecasting—Reasons, procedures, problems. NOAA Tech. Memo. NWS FCST 24, 84 pp.
Johnson, P. O., and R. W. B. Jackson, 1959: *Modern Statistical Methods: Descriptive and Inductive*. Rand McNally, 514 pp.
Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.
Kendall, M. G., and A. Stuart, 1961: *Inference and Relationship*. Vol. 2, *The Advanced Theory of Statistics*, Hafner, 676 pp.
Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperature during winter. *J. Meteor.,* **16,** 672–682.
Klugh, H. E., 1974: *Statistics: The Essentials for Research*. John Wiley and Sons, 426 pp.
Liljas, E., and A. H. Murphy, 1994: Anders Angstrom and his early papers on probability forecasting and the use/value of weather forecasts. *Bull. Amer. Meteor. Soc.,* **75,** 1227–1236.
Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Statistical Forecasting Project, Massachusetts Institute of Technology Scientific Rep. 1, 49 pp.
Marzban, C., and G. J. Stumpf, 1996: A neural network for tornado prediction based on doppler radar-derived attributes. *J. Appl. Meteor.,* **35,** 617–626.
Mason, I., 1980: Decision-theoretic evaluation of probabilistic predictions (using the relative operating characteristic). *Proc. WMO Symp. on Probabilistic and Statistical Methods in Weather Forecasting,* Nice, France, WMO, 219–228.
——, 1982a: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.
——, 1982b: On scores for yes/no forecasts. Preprints, *Ninth Conf. on Weather and Forecasting and Analysis,* Seattle, WA, Amer. Meteor. Soc., 169–173.
Miller, D. W., and M. K. Starr, 1960: *Executive Decisions and Operations Research*. Prentice Hall.
Mode, E. B., 1951: *Elements of Statistics*. Prentice-Hall, 377 pp.
Mood, A. M., 1950: *Introduction to the Theory of Statistics*. McGraw-Hill, 433 pp.
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.
——, and E. P. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.
Neter, J., and W. Wasserman, 1974: *Applied Linear Statistical Models*. Richard D. Irwin, 842 pp.
Panofsky, H. A., and G. W. Brier, 1958: *Some Applications of Statistics to Meteorology*. The Pennsylvania State University, 224 pp.
Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–1000.
Teweles, S., and W. G. Wobus, 1954: Verification of prognostic charts. *Bull. Amer. Meteor. Soc.,* **35,** 455–463.
Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. *Mon. Wea. Rev.,* **78,** 113–124.
——, 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.,* **33,** 223–226.
——, and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.,* **83,** 249–254.
Underwood, B. J., C. P. Duncan, J. A. Taylor, and J. W. Cotton, 1954: *Elementary Statistics*. Appleton-Century-Crofts, 29 pp.
White, R. M., and A. M. Galligan, 1956: The comparative accuracy of certain statistical and synoptic forecasting techniques. *Bull. Amer. Meteor. Soc.,* **37,** 1–7.
Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.