# FORECASTERS' FORUM

# Comments on "Discussion of Verification Concepts in Forecast Verification: A Practitioner's Guide in Atmospheric Science"

IAN T. JOLLIFFE AND DAVID B. STEPHENSON

*Department of Meteorology, University of Reading, Reading, United Kingdom*

(Manuscript received 22 December 2004, in final form 10 March 2005)

## 1. Introduction

We congratulate Bob Glahn on his perceptive and thoughtful review (Glahn 2004; hereafter G04) of the book we edited entitled *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (Jolliffe and Stephenson 2003; hereafter JS03). His comments will undoubtedly lead to an improved second edition. Furthermore, he has raised several very stimulating and important verification and forecasting issues that could benefit from a wider debate. We, therefore, wish to take this opportunity to respond to some of the issues raised in Glahn (2004) in the hope that a thought-provoking verification debate appears in the literature. Rather than attempt to elicit and then present a consensus response that reflects the views of all our authors (if such a thing were ever achievable!), we prefer to respond more directly to G04 with our own subjective editorial opinions. We hope that some of our authors will comment separately.

Forecast verification is an essential part of atmospheric sciences. It is the way in which the science of meteorology is ultimately judged—by the skill of its predictions. Forecast verification is an intellectually stimulating and multidisciplinary area of research that requires careful summary and interpretation of pairs of past forecasts and observations. Despite its importance, forecast verification is not always fully acknowledged in operational forecasting centers and is often completely absent from atmospheric science courses. In addition to skill, forecasts should also provide information that helps forecast users make better decisions despite the uncertainty inherent in the forecasts. These and other factors have led to more than a century of fascinating ongoing developments in forecast verification.

We agree with many of the points raised in G04 but we wish to expand here on several that we consider to be the most interesting and important issues.

## 2. Verification from a forecast developer's viewpoint

On p. 770, G04 states that "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." In other words, our book is biased toward verification for the purposes of the forecast developer rather than for the purposes of the forecast user. For example, as G04 quite rightly points out, throughout our book it is often assumed that poorly calibrated forecasts can easily be recalibrated, yet this is not always possible especially if one is a forecast user. Most forecast users would not even think of recalibrating the forecasts since they generally take the forecasts at face value; they quite naturally assume that the given forecasts are well calibrated. Unfortunately, many weather and climate forecasts are often not well calibrated (see below) and so great care needs to be exercised in judging and using such products. Furthermore, suppose a (sceptical) forecast user did want to recalibrate forecasts before verification, then he or she would often not be able to do so because of generally having insufficient access to past observations and/or knowledge of changes in the forecasting system. Our emphasis on *model-oriented* rather than *user-oriented* verification in part stems from our choice of authors for the chapters, many of whom are *verification practitioners* employed at national weather forecasting services

*Corresponding author address:* Prof. Ian T. Jolliffe, Dept. of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: i.t.jolliffe@reading.ac.uk

around the world. However, it should be noted that forecast users are generally more interested in judging how much added value forecasts can bring to their specific decision-making processes, and so are often more concerned with the assessment of user-specific forecast *value* (*utility*) rather than overall forecast *quality*. To take an extreme view, if, paradoxically, there is no guarantee that skillful forecasts will provide value to a given user, then why should any users be interested in the assessment of forecast quality?

As pointed out in chapter 1 of our book, our focus was primarily on methods for the assessment of forecast quality (forecast verification) rather than the assessment of utility, which has been addressed elsewhere (e.g., Katz and Murphy 1997). Nonetheless, we agree with Glahn that it would be good to see more user-oriented approaches to forecast verification in the literature and we hope that such approaches will be developed in the future. A more user-oriented approach to verification will help minimize some of the potential conflicts of interest caused by forecast providers assessing the quality of their own forecast products.

## 3. Resolution, reliability, and ROC for poorly calibrated systems

Calibration is a topic of fundamental importance in verification. There are basically two reasons why forecasts do not match observations:

- they are unable to discriminate between different observed situations, and
- they are poorly labeled, for example, the forecasts are on average 5°C too warm.

The ability of forecasts to discriminate between observed situations is known as *resolution,* and its existence is a necessary yet not sufficient condition for forecasts to have skill. Forecast accuracy also depends on the *reliability* (i.e., good labeling—*calibration*) of the forecasts. However, unlike resolution, reliability can be improved, *in principle,* by recalibration of the forecasts using past information about pairs of forecasts and observations. In other words, resolution is a necessary condition for skill whereas reliability is not. If forecasts have poor resolution, there is not much one can do to improve them whereas if they have poor reliability there is still hope.

On p. 770, G04 points out that presenting only the relative operating characteristic (ROC) quantities of *hit rate* ($H$) and *false alarm rate* ($F$) has "a major deficiency—it does not consider calibration." This is true since hit rate and false alarm rate are both conditional probabilities and so do not by themselves contain any information about marginal probabilities that can be used to estimate forecast frequency bias. It is easy to show that the frequency bias of binary forecasts is given by $B = H + [(1 - p)/p]F$, where $p$ is the probability of the observed event to occur (the *base rate*) and so requires knowledge of the base rate $p$ as well as ROC quantities $H$ and $F$. This helps to resolve G04's remark in the sixth paragraph of p. 772: "It is not clear to me how reliability (calibration), which is generally ignored by ROC, cannot be crucial in determining the actual economic value of forecasts." As explained in chapter 8 of JS03, the economic value is not simply a function of $H$ and $F$ but also strongly depends on the base rate $p$. Diagnostics based on ROC quantities such as the area under the ROC curve $H(F)$ are useful because they focus attention on resolution rather than reliability of the forecasts but they require careful interpretation (see Göber et al. 2004). The major deficiency in ROC not considering calibration is also a major strength—in much the same way that the product moment correlation coefficient does not measure bias but is nevertheless a useful measure of linear association for continuous forecasts. It should also be noted that, unlike ROC, economic value measures of performance have a major deficiency in that they generally have a strong dependence on the base rate and so are extremely sensitive to how forecasts are calibrated. This leads to the undesirable, yet rarely mentioned, property that economic value measures can usually be improved by hedging the forecasts.

## 4. Who should do the calibration and how should it be done?

As pointed out by G04, there is a big difference between recalibration *in principle* and what is possible *in practice.* To be able to recalibrate, one needs to have access to a suitably large sample of past pairs of forecasts and observations (not often issued to the forecast user!) and one must make certain assumptions about past and future stationarity of forecast–observation relationships in order to be able to develop a regression model suitable for performing the recalibration (such as that used in operational postprocessing schemes such as model output statistics). Perhaps more importantly, one also needs the motivation to embark upon calibration. It is often not clear who should be doing the recalibration. For example, should it be the forecast providers or should it be the forecast users themselves? It might seem obvious that it should be the forecast provider who ensures that the forecasts are well calibrated. However, it can also be argued that each user has more detailed knowledge of their particular needs and so can calibrate more optimally for their own area of application. For example, one user may be more interested in extreme temperatures in the tail of the distribution whereas another user may be more interested in more central temperatures; the calibration could then be tailored to these specific applications. Statistical postprocessing of weather and climate model predictions is an essential step in the forecasting processing that deserves to be more widely recognized. Statistical postprocessing is essential for mapping predictions made in model state space back into forecasts of real-world observations. An elegant duality between statistical postprocessing and data assimilation has recently been discovered by Stephenson et al. (2005) in the context of their work on multimodel forecast combination (Coelho et al. 2004a,b). This has led Stephenson et al. (2005) to refer to forecast postprocessing by the more dignified and meaningful expression *forecast assimilation*—the process whereby model predictions are assimilated into existing knowledge to produce improved forecasts of real-world observable quantities.

## 5. Ensemble forecasts are not the only way of making probability forecasts

Bob Glahn's "biggest disappointment" with our book is our "rolling of the verification of probability forecasts into a chapter shared by ensemble forecasting" (G04, p. 774, second paragraph). In hindsight, we agree that such a distinction would have been advantageous. We also fully agree with G04 that ensembles are only one possible way of making probability forecasts. Deterministic forecasts can easily be converted into probability forecasts by incorporating knowledge of past forecast errors. As pointed out in the glossary of JS03, deterministic (nonprobabilistic) forecasts should be considered as those for which forecast uncertainty is not provided as opposed to thinking of them as probability forecasts with zero uncertainty! In other words, deterministic forecasts should not be interpreted as *definite* or *definitive* forecasts but should be interpreted instead as *incomplete* or *poorly specified* forecasts. G04 (p. 774, sixth paragraph) notes an inconsistency of interpretation in JS03, which needs to be addressed in future editions.

As explained by G04, ensemble forecasting is a method for producing probability forecasts rather than an area of verification. This raises an interesting debate concerning ensemble forecasts: should one use ensembles to infer probabilities (using statistical postprocessing) and then verify the probability forecasts, or should one design new techniques specifically for the verification of ensemble forecasts? It is clear that in recent years, research effort has gone into designing verification methods specific for ensemble forecasts, for example, the rank histogram (Anderson 1996), multidimensional scaling (Stephenson and Doblas-Reyes 2000), bounding boxes (Weisheimer et al. 2005), the minimum spanning tree (Wilks 2004; Smith and Hansen 2004), etc. The need for forecast developers to have tools to explore their ensemble forecasts was evident in many of the talks presented at a recent World Meteorological Organization (WMO) workshop entirely devoted to ensemble forecasting (most of the talks given at the workshop are available online at http://cccma-meetings.seos.uvic.ca/ensemble/). There is, however, a growing recognition that one needs to develop probability models capable of assimilating multimodel forecast data to produce probabilities of future observable events. After all, few users explicitly request a set of multimodel ensemble predictions, and there are no users who actually live inside a model grid box!

## 6. Verification of spatial fields

The spatial nature of many meteorological forecasts (e.g., precipitation maps) poses an exciting yet difficult challenge for forecast verification. In recent years, this has become a rapidly developing area of research, which helps to partly excuse the "certain incompleteness" (G04, p. 774, paragraph 1) in chapter 6 of JS03. It is particularly difficult to summarize areas of research that are undergoing rapid development. Many diverse methods are currently being developed to tackle the verification of spatial fields of complex variables such as precipitation—for example, the intensity-scale wavelet approach recently developed by Casati et al. (2004).

A whole day was devoted to this subject at the recent WMO International Verification Methods Workshop (15–17 September 2004, held in Montreal, Quebec, Canada; electronic copies of the talks can be downloaded online from http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Workshop2004/MeetingProgram.html; see, in particular, presentation 4.1, a review by Brown and Ebert). This is an area of verification research that could benefit enormously from approaches already developed in other areas of science (e.g., medical imaging, image processing, etc.) and it is likely to undergo substantial development that we hope to describe in future editions of our book.

## 7. Statistical inference and other literature

We fully agree that the "very important topics" of sampling error, artificial skill, significance testing and, indeed, statistical inference more generally deserve more attention (G04, p. 773, paragraph 6) and this will be addressed in any future editions. The lack of emphasis in the current edition reflects that seen in the atmospheric science literature, as compared to other literature sources such as medicine; see, for example, Pepe (2003). It is hoped that this situation can be remedied in the future (see presentation 1.1, by Jolliffe, in the Montreal workshop mentioned above). A more general point is that verification, often under other names, is an important topic in several other discipline's literature, and there is much to be learned from interaction between the various disciplines involved. For example, a great deal of sophisticated research on ROC curves can be found in the medical, psychological, and signal processing literature.

## 8. Baselines and reference forecasts

On p. 770, G04 discusses possible baselines when assessing the quality or value of a forecast. The use of a well-established objective method such as model output statistics (MOS) is advocated. Certainly when assessing a new forecasting system, the main interest is likely to be a comparison with well-established methods, and such methods can be viewed as baselines in this sense. However, when constructing a skill score, the baseline or reference forecast should be one that is unskillful. Forecasts that are always the same or are chosen randomly (most varieties of climatological forecast fall into one of these categories) are suitable, but well-established forecasts that have skill are not. The position of persistence forecasts is open to debate. Since in many circumstances persistence forecasts have skill, we believe that they are not usually appropriate reference forecasts when constructing skill scores, though one would clearly want any operational forecasting system to beat them.

## 9. Terminology

A number of the comments in G04 refer to terminology. It is highly desirable to use consistent terminology and notation, but this will not always be achievable, given that many ideas have been reinvented and renamed in different disciplines. We add our own views on some of the points raised by G04.

### *a. Training data*

This terminology is not liked by G04 (p. 771, paragraph 3). We have no objection to it. "Development data" is a reasonable alternative, but has no advantage, as "training data" is well established. We feel that, as in regression, the words "dependent" and "independent" are best avoided, because of their ambiguous meaning.

### *b. Predictand*

Contrary to G04 (p. 771, paragraph 4), we rather like this word, as meaning "something that is predicted." We would encourage its use in this context, although G04 notes an inconsistency within JS03. Interestingly, although we have used it for some years, and it is not uncommon in statistical texts, it does not have an entry in the statistical dictionary by Dodge (2003).

### *c. Deterministic forecasts*

We used this in place of the ambiguous "categorical forecasts" but, like G04 (p. 771, paragraph 5), are not entirely comfortable with it. G04's suggestions of "definite" or "definitive" do not seem quite right either, but less ambiguous alternatives such as "nonprobabilistic point forecasts" or "unknown uncertainty forecasts" are rather clumsy expressions.

### *d. Divisor $(n - 1)$ or $n$ in the sample variance*

We strongly disagree with G04's *general* preference for $n$. There are two circumstances in which $n$ may be appropriate, namely when our data consist of the whole population or when the population mean $\mu$ is known and replaces the sample mean in the sum of squares. Otherwise, whenever the data are a sample from some (real or hypothetical) population and the mean of that population is unknown, there are a number of reasons to prefer $(n - 1)$. This is by far the most common situation, so it makes sense to *define the sample variance* with divisor $(n - 1)$. One theoretical reason for preferring $(n - 1)$ in this situation is that it gives an unbiased estimator of the underlying population variance, but a more compelling reason is practical rather than theoretical. Various procedures in statistical inference, such as the Student's $t$ test, have formulas that assume the sample variance is computed with divisor $(n - 1)$, the degrees of freedom associated with the estimator. Anyone wishing to conduct inferences based on a sample variance with divisor $n$ needs to use different formulas from those in the vast majority of textbooks, a recipe for error and confusion. Further discussion of the choice between $n$ and $(n - 1)$ can be found in Problem 3A.1 Note (8) of Bassett et al. (2000).

### *e. Climatology*

Despite its widespread use as a contraction of "climatological forecast," we share G04's (footnote p. 770) distaste for this terminology. However, it should be noted that "climatological forecast" is ambiguous since there are many ways one can use past climatological observations to make forecasts. For example, for continuous data forecasting the climatological mean is one possibility, but when the data are also skewed, the climatological median or mode are plausible alternatives.

## 10. Final remarks

G04 noted a number of inconsistencies in JS03, though thankfully not too many. Only two of these have been mentioned above, but we shall certainly attempt to remedy all those that were identified, in any future editions. We hope for further suggestions from the discussion that we expect to arise from G04, and we thank Bob Glahn once again for his penetrating review.

### REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Bassett, E. E., J. M. Bremner, I. T. Jolliffe, B. Jones, B. J. T. Morgan, and P. M. North, 2000: *Statistics—Problems and Solution.* 2d ed. World Scientific, 227 pp.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.,* **11,** 141–154.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson, 2004a: Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate,* **17,** 1504–1516.

——, ——, ——, ——, and ——, 2004b: Skill of coupled model seasonal forecasts: A Bayesian assessment of ECMWF ENSO forecasts. ECMWF Tech. Memo. 426, 16 pp.

Dodge, Y., 2003: *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 498 pp.

Glahn, H. R., 2004: Discussion of verification concepts in *Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wea. Forecasting,* **19,** 769–775.

Göber, M., C. A. Wilson, S. F. Milton, and D. B. Stephenson, 2004: Fairplay in the verification of operational quantitative precipitation forecasts. *J. Hydrol.,* **288,** 225–236.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.

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

Pepe, M. S., 2003: *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, 302 pp.

Smith, L. A., and J. A. Hansen, 2004: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Wea. Rev.,* **132,** 1522–1528.

Stephenson, D. B., and F. J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus,* **52A,** 300–322.

——, C. A. S. Coelho, M. Balmaseda, and F. J. Doblas-Reyes, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus,* **57A,** 253–264.

Weisheimer, A., L. A. Smith, and K. Judd, 2005: A new view of seasonal forecast skill: Bounding boxes from the DEMETER ensemble forecasts. *Tellus,* **57A,** 265–279.

Wilks, D. S., 2004: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.,* **132,** 1329–1340.