NOTES AND CORRESPONDENCE

Evaluating Probabilistic Forecasts Using Information Theory

MARK S. ROULSTON AND LEONARD A. SMITH*

Pembroke College, Oxford University, Oxford, United Kingdom

26 January 2001 and 6 November 2001

ABSTRACT

The problem of assessing the quality of an operational forecasting system that produces probabilistic forecasts is addressed using information theory. A measure of the quality of the forecasting scheme, based on the amount of a data compression it allows, is outlined. This measure, called ignorance, is a logarithmic scoring rule that is a modified version of relative entropy and can be calculated for real forecasts and realizations. It is equivalent to the expected returns that would be obtained by placing bets proportional to the forecast probabilities. Like the cost–loss score, ignorance is not equivalent to the Brier score, but, unlike cost–loss scores, ignorance easily generalizes beyond binary decision scenarios. The use of the skill score is illustrated by evaluating the ECMWF ensemble forecasts for temperature at London's Heathrow airport.

1. Introduction

Operational weather forecasters now recognize that uncertainties in the initial conditions used to initialize numerical weather prediction (NWP) models, as well as errors in the models themselves, lead to uncertainty in the forecast. Many forecast centers now attempt to estimate the impact of these uncertainties in the initial conditions by generating ensembles of forecasts (Molteni et al. 1996; Toth and Kalnay 1997). The ensemble forecast members usually differ in initial conditions, although research is under way into generating ensembles that reflect model error (Houtekamer et al. 1996; Buizza et al. 1999; Stensrud et al. 1999; Evans et al. 2000), and such methods are now becoming operational. The best method to construct these ensembles is the subject of current research (for a review, see Palmer 2000). Ensemble forecasting is a Monte Carlo approach to sampling the forecast probability distribution function (PDF); explicit calculation of this function in the state space of a modern NWP model is computationally impossible, and possibly ill defined (Smith et al. 1999). Computational limits determine the size of the ensembles generated.

Users of weather forecasts may benefit significantly from the greater amount of information contained in a probabilistic forecast than in a single deterministic forecast (Smith et al. 2001; Richardson 2000, 2001; Roulston and Smith 2002). To ascertain this benefit to a particular user one should create a cost function that takes into account the decisions that the user can make, and the utility associated with possible outcomes. The result will be specific to that particular user.

The question of how to assess the general quality of probabilistic forecasts is a subject of current research in the weather forecasting community. Currently used methods include the Brier score (Brier 1950), ranked probability score (Epstein 1969; Murphy 1971), relative operating characteristics (Swets 1973; Mason 1982), rank histograms (Anderson 1996; Hamill and Colucci 1996; Talagrand et al. 1997), and the generalization of rank histograms to higher dimensions (Smith 2000).

Information theory provides a useful theoretic framework to understand and quantify weather and climate predictability (Leung and North 1990; Schneider and Griffies 1999; Kleeman 2002). It was suggested by Leung and North (1990) that a relative entropy type measure might be used as the basis of a skill score for deterministic forecasts. Information theoretic measures, such as entropy, have been used in previous studies to quantify ensemble spread (Stephenson and Doblas-Reyes 2000). In these studies the entropy of the probabilistic forecast was suggested as a predictor of forecast skill, rather than as a measure of forecast skill.

In this paper, we propose a wider role for a logarithmic scoring rule (Lindley 1985) by showing how it fits

*Additional affiliation: Centre for the Analysis of Time Series, London School of Economics, London, United Kingdom.

Corresponding author address: Mark S. Roulston, Pembroke College, St. Aldates, Oxford OX1 1DW, United Kingdom.
E-mail: roulston@maths.ox.ac.uk

into the context of information theory. Under this scoring the ''best'' forecast would be one that leads to the highest level of data compression when describing truth; this forecast would also yield the highest expected return if used to place proportional bets on the future. The idea of using data compressibility as a measure of model quality has philosophical appeal (Davies 1991), while the correspondence with gambling returns has some relevance to insurance and weather derivative pricing applications, or any other industry with the option to take action based on a forecast.

2. Ignorance defined

The aim is to develop a forecast skill score that measures the quality of the forecast PDF. The forecast PDF should be assessed on how similar it is to the true PDF. In this paper, the phrase ''true PDF'' means the PDF of consistent initial conditions evolved forward in time under the dynamics of the real atmosphere (Smith et al. 1999). This initial PDF is the product of the distribution of observational uncertainty and the distribution of states on the atmospheric attractor (if one exists).

Consider two PDFs, defined by the vectors **p** and **f**. Let the ith component of these vectors define the probability of the ith outcome occuring; hence, $\Sigma\, p_i = \Sigma\, f_i = 1$. One measure of ''distance'' between **p** and **f** is the relative entropy given by

$$D(\mathbf{p}\,|\,\mathbf{f}) = \sum_i (p_i \log_2 p_i - p_i \log_2 f_i). \tag{1}$$

Relative entropy is not a true distance; it satisfies neither the requirement of symmetry nor the triangle inequality (Cover and Thomas 1991). A scoring rule, based on optimal data compression and closely related to relative entropy, will now be described.

Classify every event into one, and only one, of n possible outcomes. A model generates a probabilistic forecast of the outcome of any event: the probability of the ith event according to the probabilistic prediction system is f_i (where $i = 1, \ldots, n$). Before the event occurs, a data compression scheme to encode the actual outcome is designed. The simplest encoding scheme would assign $\log_2 n$ bits to each outcome, since this is the number of bits required to encode n integers. This encoding scheme, however, would not be the most efficient. Greater compression can be achieved by assigning fewer bits to the most likely outcomes and more bits to the less likely outcomes. A fundamental result of information theory says that, if the probabilities of the n outcomes are given by the f_i, then the optimal data compression scheme assigns B_i bits to outcome i, where B_i is given by (Shannon 1948)

$$B_i = -\log_2 f_i. \tag{2}$$

The details of the encoding scheme are not important for this argument; the existence of such a scheme merely provides a philosophical basis for the skill score.

Note that if $f_i = 0$ then, according to Eq. (2), an infinite number of bits is assigned to the ith outcome. This is because an optimal compression scheme would have no way of encoding any outcome deemed impossible a priori. This raises the interesting issue of whether reporting 0 forecast probabilities can ever be justified, especially if the forecast probabilities are estimates obtained from finite ensembles and imperfect models. Forecasters should replace 0 forecast probabilities with small probabilities based on the uncertainties in the forecast PDF. Not to do so means reporting the improbable as the impossible. This would violate ''Cromwell's rule,''[1] which warns against assigning 0 probability to an event unless it is truly impossible (Lindley 1985).

The information-based ignorance score has a simple interpretation. Suppose that person A and person B are both in possession of the probabilistic forecast defined by f_i ($i = 1, \ldots, n$). Person A knows what the actual outcome is and he is going to send a message to person B, telling her this outcome. They have agreed to use an optimal data encoding scheme defined by the f_i. How many bits must A send to B? If the actual outcome is j, then the number of bits that A must send is IGN $= -\log_2 f_j$. This, therefore, is the information deficit, or ignorance, of person B when she had the probabilistic forecast but before A sent her the message telling her the actual outcome. This value of ignorance is for one forecast and realization; thus, it is a scoring rule (Murphy 1997). It should be averaged over a verification dataset of T forecast–realization pairs. That is,

$$\langle \mathrm{IGN} \rangle = -\frac{1}{T} \sum_{k=1}^{N} \log_2 f(k)_{j(k)}, \tag{3}$$

where $f(k)_i$ is the probability of outcome i according to the probabilistic forecast k and $j(k)$ is the corresponding actual outcome. This is the number of bits that A must send to B to describe the true outcome, averaged over all the verification forecasts.

Let the true PDF be represented by p_i. This PDF is generally unknown. The expected value of the ignorance, E[IGN], of a particular forecast is given by

$$E[\mathrm{IGN}] = -\sum_{i=1}^{n} p_i \log_2 f_i. \tag{4}$$

The relative entropy of the true and forecast PDFs can thus be written as

$$D(\mathbf{p}\,|\,\mathbf{f}) = E[\mathrm{IGN}] - H(\mathbf{p}), \tag{5}$$

where $H(\mathbf{p})$ is the entropy of the true PDF given by

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{6}$$

As stated, the true PDF will be unknown so the $H(\mathbf{p})$

[1] ''I beseech you, in the bowels of Christ, think it possible you may be mistaken.'' (Oliver Cromwell in a letter to the General Assembly of the Church of Scotland, 3 August 1650.)

term in Eq. (5) cannot be calculated. Therefore, there is no way to know when a perfect model [$D(\mathbf{p} \mid \mathbf{f}) = 0$] has been obtained, although since $D(\mathbf{p} \mid \mathbf{f})$ cannot be negative no model will have a lower value of E[IGN] than a perfect model. It can be shown that E[IGN] has a single minimum at $f_i = p_i$; this is equivalent to the statement that H is the minimum number of bits required to describe the data (Shannon 1948). Furthermore, if a forecaster's best estimate of p_i is f_i, and the forecast they intend to issue is g_i, then the predicted expected ignorance of the forecast is

$$\text{Pred}(E[\text{IGN}]) = -\sum_{i=1}^{n} f_i \log_2 g_i. \tag{7}$$

The minimum value of Pred (E[IGN]) occurs only when $g_i = f_i$; thus, ignorance is a strictly proper scoring rule (Lindley 1985; Murphy and Daan 1985; Winkler and Murphy 1968): forecasters cannot expect to reduce their ignorance score by issuing a PDF different from their best judgment. The predicted expected ignorance when $g_i = f_i$ is thus

$$\text{Pred}(E[\text{IGN}]) = -\sum_{i=1}^{n} f_i \log_2 f_i = H(\mathbf{f}). \tag{8}$$

This quantity is the entropy of the forecast, which has been suggested as a predictor of forecast skill (Stephenson and Doblas-Reyes 2000). Equation (8) shows the relationship between $H(\mathbf{f})$ and predicted skill, as measured by ignorance explicitly. It suggests that, averaged over many forecasts, the ignorance should be the same as the average entropy of the forecasts. Satisfying this consistency condition, however, does not imply that the forecasts have the minimum possible ignorance: climatological forecasts would satisfy the condition.

3. Relationship between ignorance and forecast quality

The ignorance score makes no assumptions about the shape of the PDF. Nevertheless, to illustrate some of the properties of the ignorance score it will be assumed that the variable in question, x, is continuous and that both the true PDF ρ_{truth} and the forecast PDF ρ_{fcst} are normal distributions. That is, ρ_{truth} and ρ_{fcst} are defined by

$$\rho_{\text{truth}}(x) = \frac{1}{\sigma_{\text{truth}}\sqrt{2\pi}} \exp\left[-\frac{(x - \overline{x}_{\text{truth}})^2}{2\sigma_{\text{truth}}^2}\right] \quad \text{and} \tag{9}$$

$$\rho_{\text{fcst}}(x) = \frac{1}{\sigma_{\text{fcst}}\sqrt{2\pi}} \exp\left[-\frac{(x - \overline{x}_{\text{fcst}})^2}{2\sigma_{\text{fcst}}^2}\right]. \tag{10}$$

The expected ignorance can thus be calculated:

$$E[\text{IGN}] = -\frac{1}{\ln 2}\int_{-\infty}^{+\infty} \rho_{\text{truth}}(x) \ln \rho_{\text{fcst}}(x)\, dx$$

$$= \frac{1}{2\ln 2}\left[\ln 2\pi + \ln\sigma_{\text{fcst}}^2 + \frac{\sigma_{\text{truth}}^2 + (\overline{x}_{\text{truth}} - \overline{x}_{\text{fcst}})^2}{\sigma_{\text{fcst}}^2}\right]. \tag{11}$$

Using Eq. (11), it can be seen how the ignorance is affected by conventional aspects of forecast quality (Murphy 1997). Bias in the forecast, $(\overline{x}_{\text{truth}} - \overline{x}_{\text{fcst}})$, causes an increase in E[IGN]. Greater uncertainty of reality, σ_{truth}, also leads to greater expected ignorance. The effect of sharpness, σ_{fcst}, on ignorance is not monotonic; in particular, there is a unique minimum in E[IGN] when $\sigma_{\text{fcst}}^2 = \sigma_{\text{truth}}^2 + (\overline{x}_{\text{truth}} - \overline{x}_{\text{fcst}})^2$. Thus for an unbiased forecast the minimum ignorance is obtained when the variance of the forecast equals the variance of the perfect forecast PDF, but if the forecast is biased then the ignorance is minimized by increasing the forecast variance to partially compensate for the bias. Note that a forecast with a smaller bias will still have a lower ignorance for *any* given value of σ_{fcst}. The ignorance measures reliability in that it is a minimum if, and only if, truth is picked from the forecast PDF. The skill of the forecast, that is, its accuracy relative to other forecasts, can be calculated by simply calculating the difference of the ignorance of the forecasts that are being compared. This is because ignorance is an information and information is an additive quantity. Each bit of ignorance represents a factor-of-2 increase in uncertainty.

4. Relationship between ignorance and cost–loss

There is a direct correspondence between data compression and gambling returns (Kelly 1956; Cover and Thomas 1991). If a gambler can bet an arbitrary fraction w_i of their wealth on outcome i ($\Sigma_i\, w_i = 1$) then, to maximize their expected return averaged over sequential bets, gamblers should bet proportionally, that is, bet a fraction f_i of their wealth on the ith outcome occurring. If this strategy is adopted, the ratio of the gamblers' wealth after the bet to that before the bet has an expected value of 2^W, where

$$W = \sum_{i=1}^{n} p_i \log_2 o_i w_i$$

$$= \sum_{i=1}^{n} p_i \log_2 w_i + \sum_{i=1}^{n} w_i \log_2 o_i \tag{12}$$

and o_i is the odds (wealth multiplier) assigned to outcome i. To maximize the expected return, averaged over sequential bets, gamblers should bet proportionally and set $w_i = f_i$. If the house sets odds based on a forecast probability distribution g_i by setting $o_i = 1/g_i$, then Eq. (12) becomes

$$W = E[\text{IGN}]_{\text{house}} - E[\text{IGN}]_{\text{gambler}}. \tag{13}$$

Thus, gamblers can only expect to make money if they have lower ignorance than the house.

When considering the economic value of forecasts, the binary cost–loss scenario is commonly used (Katz and Murphy 1987, 1997). In this scenario there are two outcomes (e.g., not freezing and freezing). The user can make a decision to protect or not to protect (e.g., to grit

the roads or not to grit the roads). This protection has a cost C but, should the user choose not to protect and adverse weather occurs, the user sustains a loss L. Let the probability of it freezing be $p_1 = p$, and the probability of it not freezing be $p_2 = 1 - p$. The wealth multipliers o_i associated with each outcome are $o_1 = L/C - 1$ and $o_2 = 1$. However, in the simple cost–loss scenario, the users cannot spread their wealth arbitrarily between the outcomes. Since the potential loss the users can suffer is L, this is the amount of wealth they can bet on the outcomes. Effectively they must either bet $w_1 = 0$, $w_2 = 1$ or $w_1 = C/L$, $w_2 = 1 - C/L$. They would choose the latter if p is greater than C/L. If p is less than C/L, the user could replicate the proportional betting strategy (by gritting a fraction pL/C of the roads, if this is possible). The cost–lost score is parametric; it depends on the value of C/L. If a uniform distribution of C/L ratios is assumed, it can be shown that the mean cost–loss score is equivalent to the Brier score (Murphy 1966; Richardson 2001). The advantage of ignorance over the cost–lost score is that ignorance easily generalizes beyond the binary decision case; indeed ignorance can be defined for a continous distribution $\rho(x)$ as IGN $= -\log_2\rho(x_a)$, where x_a is the actual outcome. It can be shown that ignorance is equivalent to the cost–loss score averaged over a distribution of cost–loss ratios that is weighted toward values of C/L close to 0 and unity (see the appendix).

5. Relationship between ignorance and Brier score

The Brier score is a common skill score for assessing probabilistic forecasts (Brier 1950). It will now be shown that a forecast scheme with a lower expected Brier score than another forecast scheme may not necessarily have a lower value of expected ignorance. In the simple two-outcome case, ignorance is a double-valued function of Brier score. It shares this property with the cost–loss value for a single coss–loss ratio (Murphy and Ehrendorfer 1987).

Consider an event with n possible outcomes. Let f_i be the forecast probability of the ith outcome. Let j be the actual outcome. The Brier score BS is given by

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - \delta_{ij})^2, \tag{14}$$

where $\delta_{ij} = 0$ when $i \neq j$ and $\delta_{ij} = 1$ when $i = j$. Thus the Brier score of forecast f_i if the outcome is j is

$$\mathrm{BS} = \frac{1}{n}\left(\sum_{i=1}^{n} f_i^2 - 2f_j + 1\right). \tag{15}$$

If the true probability distribution is p_i, then the expected value of the Brier score, E[BS], is given by

$$E[\mathrm{BS}] = \sum_{j=1}^{n} p_j \frac{1}{n}\left(\sum_{i=1}^{n} f_i^2 - 2f_j + 1\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} f_i^2 - 2\sum_{i=1}^{n} p_i f_i + 1\right). \tag{16}$$

Consider the case when there are two possible outcomes. Let the true probability distribution be given by $(p, 1 - p)$. Let the model probability distribution be given by $(f, 1 - f)$, where $f = p + \Delta$. Equation (16) gives the expected Brier score, E[BS]:

$$E[\mathrm{BS}] = (1/2)[(p + \Delta)^2 + (1 - p - \Delta)^2 - 2p(p + \Delta) - 2(1 - p)(1 - p - \Delta) + 1], \tag{17}$$

which can be simplified to give

$$E[\mathrm{BS}] = \Delta^2 - p^2 + p. \tag{18}$$

From Eq. (18) it can be seen that there are two models that have the same expected Brier score. The Δ values of these models are given by

$$\Delta = \pm\sqrt{E[\mathrm{BS}] + p^2 - p} = \pm|\Delta|. \tag{19}$$

The expected ignorance of each of these models is given by

$$E[\mathrm{IGN}] = -p\log_2(p + \Delta) - (1 - p)\log_2(1 - p - \Delta). \tag{20}$$

The difference between the expected ignorance for these two models is thus

$$E[\mathrm{IGN}]_+ - E[\mathrm{IGN}]_- = p\log_2\left(\frac{p - |\Delta|}{p + |\Delta|}\right) - (1 - p)\log_2\left(\frac{1 - p - |\Delta|}{1 - p + |\Delta|}\right). \tag{21}$$

From Eq. (21) it can be seen that the two values of Δ, corresponding to a single value of the expected Brier score, give two different values of the expected ignorance. The two branches merge when $\Delta = 0$, $(f = p)$, demonstrating that a perfect model will always have both a lower expected Brier score and a lower expected ignorance than any imperfect model. If $p = 0.5$, the branches are coincident, so in this case a lower expected Brier score implies a lower expected ignorance. When $p \neq 0.5$, however, the branches are distinct. The branches correspond to f underestimating and overestimating the value of p, respectively. Thus if model A and model B lie on different branches, it is possible that model A may appear better than model B if judged by Brier score but that model B will be deemed better if ignorance is used instead. Given the relationship between ignorance and gambling returns, this result means that a house setting odds based on a minimum Brier score model will be expected to lose money to a gambler using the

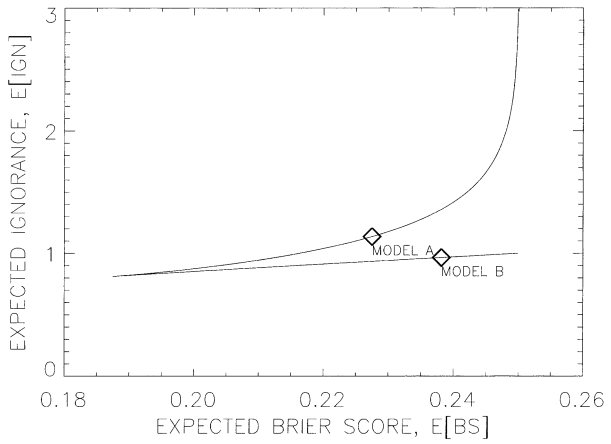

FIG. 1. A plot of expected ignorance against expected Brier score for a binary event with $p = 0.25$. The curves are parameterized by the forecast probability f. The curves intersect when $f = p$. Model A has $f = 0.05$, and model B has $f = 0.475$.

model with the lower ignorance. An example of two such models is shown in Fig. 1.

Ignorance is a double-valued function of the expected Brier score because, while the expected Brier score is symmetric in f, ignorance is asymmetric, as is the cost–loss value for a fixed cost–loss ratio.

6. Ignorance and continuous forecast variables: An example

Rank histograms are often used when the forecast variable is continuous on one dimension. Suppose there is an ensemble of N forecasts of the variable X. If the ensemble members have been picked from the true PDF, truth is equally likely to fall between any two members of the ensemble. Therefore, the forecast PDF is often approximated by a uniform distribution between each ensemble member. Let the ensemble members be ranked: X_i (where $i = 1, \ldots, N$). The forecast probability density between X_i and X_{i+1} is

$$f_i = \frac{1}{(N + 1)\Delta X_i}, \tag{22}$$

where

$$\Delta X_i = X_{i+1} - X_i \qquad 0 < i < N$$

$$\Delta X_0 = X_1 - X_{\min} \qquad \Delta X_N = X_{\max} - X_N, \tag{23}$$

where $[X_{\min}, X_{\max}]$ is the a priori interval on which X is expected to be. If truth lies in the jth interval, then the ignorance, which is the number of bits required to specify truth, is given by

$$\text{IGN} = \log_2(N + 1) + \log_2 \Delta X_j. \tag{24}$$

The first term on the rhs of Eq. (24) is the number of bits required to specify between which two ensemble members truth lies. The second term is the number of bits required to specify where in this interval truth actually lies. The expected ignorance is the expected value of the number of bits that will be required. This will clearly depend on the choice of $X_{\min}$ and $X_{\max}$ since truth will sometimes fall outside the ensemble. Calculation of ignorance from rank histograms assumes that the user requires the same resolution of forecast in the interval $[X_{\min}, X_{\max}]$. If this is not the case, categorical forecasts should be used to evaluate the ignorance score instead. The categories of a continous forecast variable can be based on climatology. They can be chosen so that each category is equally probable according to the climatology. If this is done, then, if there are n categories, the ignorance of a forecast based on climatology will be $\log_2 n$. The fractional ignorance can then be defined as

$$IGN = -\frac{\log_2 f_j}{\log_2 n}, \tag{25}$$

where f_j is the probability that the forecast system assigned to the actual outcome. If $IGN < 1$, then the forecast contains more information than the climatology. If $IGN = 1$, then the forecast is no better than a forecast based on climatology. If $IGN > 1$, then the forecast actually contains *less* information than the climatology. This can happen if the forecast is more precise (i.e., confident) but less accurate than a climatological forecast (e.g., a forecast PDF that is narrow but in the wrong place). In this situation it is not unreasonable to describe the forecast as worse than useless; the forecast could cause a user to make a decision less optimal than the decision made based on climatology alone.

7. Using ignorance: Temperature at Heathrow

To illustrate the use of the ignorance skill score, a simple example is presented. Figure 2a shows the observed temperature at London's Heathrow airport for almost two years, starting from February of 1999. The thick line is an average seasonal cycle. Consider the simple binary forecast of whether the temperature will be above or below average for the time of year. In the absence of any forecast, a person's ignorance concerning this question is 1 bit. With the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecast a probabilistic forecast can be constructed by simply counting the fractions of ensemble members that are above or below the seasonal average. An extra, fictitious, ensemble member can be split equally between the two possibilities as a simple way to account for uncertainty due to the finite ensemble size. If there are N ensemble members, n discrete outcomes (in this case $n = 2$), and the number of ensemble members with the ith outcome is Q_i, then the values of f_i, with the fictitious ensemble member included, is given by

$$f_i = \frac{Q_i + (1/n)}{N + 1}. \tag{26}$$

In Eq. (26), it can be seen that an extra ensemble mem-

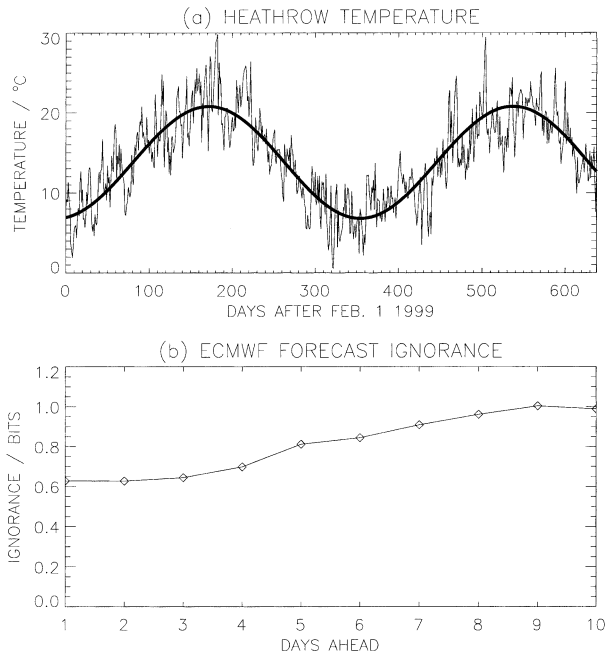

FIG. 2. (a) The observed temperature at London's Heathrow airport (thin line) and an average seasonal cycle (thick line). (b) The average ignorance of probabilistic forecasts of whether the temperature will be above or below the seasonal average. The daily forecasts were constructed using operational 51-member ECMWF ensembles.

ber has been equally distributed over all the bins. This is a crude method to account for the finite ensemble size; more sophisticated approaches that incorporate climatology are possible. Let f be the forecast probability that the temperature will be above average. If the temperature is indeed above average, the ignorance of the forecast is $-\log_2 f$. If the temperature is not above average, the ignorance is $-\log_2(1 - f)$. Figure 2b shows the ignorance for the ECMWF ensembles of temperature at Heathrow averaged over the period shown in Fig. 2a. Out to 3 days ahead, the average ignorance is about 0.63 bits. This represents a substantial improvement over the 1 bit of ignorance without a forecast. If you were offered even odds of the temperature being below average and you bet your wealth proportionally according to the forecast, you would expect to increase your wealth by 29% per bet. Not suprising, the average ignorance increases with lead time and reaches 1 bit at 9 days. This means that the 51-member ensemble forecast contains information up to day 8; beyond day 9, however, it cannot be distinguished from climatology.

8. Summary

A skill score for assessing probabilistic forecasts based on the information deficit (or ignorance) given the forecast has been presented. This skill score is directly related to the level of data compression that could be achieved using the forecast to design the compression algorithm. The relationship between data compression and gambling returns implies that this skill score corresponds to the expected returns of a gambler placing optimal (i.e., proportional) bets on the possible outcomes. The relationship of ignorance to gambling is not generally equivalent to the cost–loss score, which is used in simple studies of the economic value of forecasts. The correspondence between gambling returns and ignorance only holds if the user is free to adopt the optimal proportional (''Kelly'') betting strategy. In the cost–loss scenario this is not the case. Also, ignorance easily generalizes beyond binary decision scenarios.

The ignorance score does not indicate what effects are contributing to the loss of skill (e.g., greater ensemble spread or because truth is lying outside the ensemble). No skill score that attempts to summarize probabilistic forecast skill in a single number can describe such effects. If such a single number summary is required, however, the ignorance has advantages over other scores such as the Brier score and the cost–loss ratio.

Ignorance also has a more robust philosophical justification than the Brier score. Ignorance directly measures the average information deficit of someone in possession of a particular forecasting model. Using ignorance naturally connects the problems of practically evaluating real forecasts to the information-theoretic framework for weather and climate prediction, which has been constructed by other workers in the field (Leung and North 1990; Kleeman 2002).

The ignorance can be calculated either for categorical forecasts constructed from ensembles or from rank histograms by considering how much information is required to specify the location of truth in the ordered ensemble.

Given its advantages over other skill scores, it is likely to prove a particularly useful tool in future evaluation of probabilistic forecasts, which is a relatively neglected aspect of current meteorological research.

Acknowledgments. The authors thank the two anonymous reviewers whose suggestions greatly improved this paper. This work was supported by ONR DRI Grant N00014-99-1-0056.

APPENDIX

Skill Scores and Cost–Loss

This appendix derives the relationships between cost–loss scores and the quadratic (Brier) and logarithmic (ignorance) skill scores.

The cost–loss score is the realized loss of users attempting to minimize their expected loss. Let the users' cost–loss matrix for a binary event be

	Event happens	Event does not happen
User acts	C	C
User does not act	L	0

where C is the cost of acting, and L is the loss incurred if action is not taken and the event occurs. If the forecast probability of the event is f, then, to minimize their expected loss, the user should act if $f \geq C/L$. If the actual probability of the event is p, then the expected loss U of the user will be

$$U = \begin{cases} C & \text{when } f \geq C/L \\ pL & \text{when } f < C/L. \end{cases} \quad \text{(A1)}$$

If $u(\alpha)$ is the density of users with a cost–loss ratio of $\alpha = C/L$, then the expected loss, averaged over users and normalized in units of L, is

$$\langle U \rangle = \int_0^f \alpha u(\alpha)\, d\alpha + p \int_f^1 u(\alpha)\, d\alpha \qquad f \in [\varepsilon, 1 - \varepsilon], \quad \text{(A2)}$$

where ε is the uncertainty in f that should usually be included. Differentiation w.r.t. f gives

$$\frac{d\langle U \rangle}{df} = u(f)(f - p). \quad \text{(A3)}$$

The expected quadratic (Brier) score is given by

$$\begin{aligned} \langle \mathrm{BS} \rangle &= p[(1 - f)^2 + (1 - f)^2] \\ &\quad + (1 - p)\{f^2 + [1 - (1 - f)]^2\} \\ &= 2p - 4pf + 2f^2. \end{aligned} \quad \text{(A4)}$$

Differentiation of Eq. (A4) w.r.t. f gives

$$\frac{d\langle \mathrm{BS} \rangle}{df} = 4(f - p). \quad \text{(A5)}$$

A comparison of Eqs. (A3) and (A5) shows that the expected Brier score is linear with the average expected loss if $u(\alpha)$ is uniform (Murphy 1966; Richardson 2001).

The expected logarithmic (ignorance) score is given by

$$\langle \mathrm{IGN} \rangle = -p \log f - (1 - p) \log(1 - f). \quad \text{(A6)}$$

Differentiation gives

$$\frac{d\langle \mathrm{IGN} \rangle}{df} = \frac{f - p}{f(1 - f)}. \quad \text{(A7)}$$

Comparing Eqs. (A3) and (A7) indicates that the expected ignorance is linear with the average expected loss if $u(\alpha) \propto [\alpha(1 - \alpha)]^{-1}$, where $\alpha \in [\varepsilon, 1 - \varepsilon]$. So, the ignorance score is linear with the overall cost–loss value for a distribution of users heavily weighted at cost–loss ratios close to 0 and unity. The distribution is singular at $\alpha = 0$ and $\alpha = 1$. At these values, no decision-making scenario exists, since a user with $\alpha = 0$ would always act and a user with $\alpha = 1$ would never act.

REFERENCES

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

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

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125,** 2887–2908.

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory.* John Wiley, 542 pp.

Davies, P. C. W., 1991: Why is the physical world so comprehensible? *Complexity, Entropy and the Physics of Information,* W. H. Zurek, Ed., Addison-Wesley, 61–70.

Epstein, E., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8,** 985–987.

Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from The Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128,** 3104–3127.

Hamill, T. M., and S. J. Colucci, 1996: Random and systematic error in NMC's short-range Eta ensembles. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences,* San Francisco, CA, Amer. Meteor. Soc., 51–56.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124,** 1225–1242.

Katz, R. W., and A. H. Murphy, 1987: Quality/value relationships for imperfect information in the umbrella problem. *Amer. Stat.*, **41,** 187–189.

——, and ——, 1997: Forecast value: Prototype decision-making models. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 183–217.

Kelly, J., 1956: A new interpretation of information rate. *Bell Syst. Technol. J.*, **35,** 916–926.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, in press.

Leung, L.-Y., and G. R. North, 1990: Information theory and climate prediction. *J. Climate*, **3,** 5–14.

Lindley, D. V., 1985: *Making Decisions.* John Wiley and Sons, 207 pp.

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122,** 73–119.

Murphy, A. H., 1966: A note on the utility of probabilistic predictions and the probability score in the cost–loss ratio decision situation. *J. Appl. Meteor.*, **5,** 534–537.

——, 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10,** 155–156.

——, 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts,* A. H. Murphy and R. W. Katz, Eds., Cambridge University Press, 19–70.

——, and H. Daan, 1985: Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

——, and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Wea. Forecasting,* **2,** 243–251.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63,** 71–116.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126,** 649–667.

——, 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, in press.

Roulston, M. S., and L. A. Smith, 2002: End-to-end ensemble forecasting: Ensemble interpretation in forecasting and risk management. Preprints, *Symp. on Observations, Data Assimilation, and Probabilistic Prediction,* Orlando, FL, Amer. Meteor. Soc., 123–126.

Schneider, T., and S. M. Griffies, 1999: A conceptual framework for predictability studies. *J. Climate,* **12,** 3133–3155.

Shannon, C. E., 1948: A mathematical theory of communication. *Bell Syst. Technol. J.,* **27,** 379–423, 623–656.

Smith, L. A., 2000: Disentangling uncertainty and error: On the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics,* A. I. Mees, Ed., Birkhauser, 31–64.

——, C. Ziehmann, and K. Fraedrich, 1999: Uncertainty dynamics and predictability in chaotic systems. *Quart. J. Roy. Meteor. Soc.,* **125,** 2855–2886.

——, M. S. Roulston, and J. Hordenberg, 2001: End to end ensemble forecasting: Towards evaluating the economic value of the ensemble prediction system. ECMWF Tech. Rep. 336.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

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

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–999.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictibility,* Reading, United Kingdom, ECMWF, 1–25.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Winkler, R. L., and A. H. Murphy, 1968: ''Good'' probability assessors. *J. Appl. Meteor.,* **7,** 751–758.