Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness

James Mitchell[1] and Kenneth F. Wallis[2]

[1]National Institute of Economic and Social Research
[J.Mitchell@niesr.ac.uk]

[2]University of Warwick
[K.F.Wallis@warwick.ac.uk]

Revised, 21 November 2008

Summary. In a recent article Gneiting, Balabdaoui and Raftery (*JRSS*B, 2007) propose the criterion of sharpness for the evaluation of predictive distributions or density forecasts. They motivate their proposal by an example in which standard evaluation procedures based on probability integral transforms cannot distinguish between the ideal forecast and several competing forecasts. In this paper we show that their example has some unrealistic features from the perspective of the time-series forecasting literature, hence it is an insecure foundation for their argument that existing calibration procedures are inadequate in practice. We present an alternative, more realistic example in which relevant statistical methods, including information-based methods, provide the required discrimination between competing forecasts. We conclude that there is no need for a subsidiary criterion of sharpness.

Keywords: Calibration; Density forecast; Kullback-Leibler Information Criterion; Probability integral transform; Tests of autocorrelation; Tests of fit

JEL Classification numbers: C22, C53

Acknowledgments Some preliminary results were presented at the Oxford Forecasting and Decision Analysis Group, March 2008. We are grateful to seminar participants, Michael Clements, John Geweke and Tilmann Gneiting for comments and discussion. James Mitchell also acknowledges support by the ESRC under award RES-000-22-1390.

1. Introduction

Forecasts for an uncertain future are increasingly presented probabilistically. Tay and Wallis (2000) survey applications in macroeconomics and finance, and more than half of the inflation targeting central banks, worldwide, now present density forecasts of inflation in the form of a fan chart. When the focus of attention is the future value of a continuous random variable, the presentation of a density forecast or predictive distribution – an estimate of the probability distribution of the possible future values of the variable – represents a complete description of forecast uncertainty. It is then important to be able to assess the reliability of forecasters' statements about this uncertainty. Dawid's prequential principle is that assessments should be based on the forecast-observation pairs only; this 'has an obvious analogy with the Likelihood Principle, in asserting the irrelevance of hypothetical forecasts that might have been issued in circumstances that did not, in fact, come about' (Dawid, 1984, p.281). A standard approach is to calculate the probability integral transform values of the outcomes in the forecast distributions, and assessment rests on 'the question of whether [such] a sequence "looks like" a random sample from $U[0,1]$' (p.281; quotation marks in the original). If so, the forecasts are said to be well-calibrated.

In a recent article, Gneiting, Balabdaoui and Raftery (2007) propose the paradigm of maximising the sharpness of the predictive distributions subject to calibration for the evaluation of forecasts. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only, although the condition of calibration remains a property of the forecast-observation pairs. They motivate their argument by an example in which four different forecasts are shown, in a simulation experiment, to produce uniform probability integral transforms, hence this requirement cannot distinguish between them. Since one of them is the 'ideal' or correct forecast, 'this is a disconcerting result' (2007, p.245), which leads to the authors' argument that there is a need for additional criteria. However their example has some particular features which, from the point of view of practical time-series forecasting, make it an insecure foundation on which to base their claim that existing evaluation methods are inadequate. One such feature is the absence of a time dimension, while others concern the nature of the competing forecasts and the limited evaluation criteria employed in the example. Our first purpose in this paper is to elaborate these shortcomings, also utilising test procedures based on the Kullback-Leibler Information Criterion (KLIC). We then provide a more realistic example in which several competing

forecasts produce uniform probability integral transforms, yet this is not a 'disconcerting result' because the calibration requirement as posed by Dawid can indeed distinguish the 'ideal' forecast from its competitors in typical time-series forecasting contexts. Hence we question the perceived need for the sharpness criterion.

The rest of the paper proceeds as follows. Section 2 describes the statistical framework for the problem at hand and the evaluation methods to be employed. Section 3 presents the example of Gneiting, Balabdaoui and Raftery (2007), hereafter GBR, and describes its limitations. Section 4 presents a second example, in which available statistical methods, without an explicit sharpness criterion, satisfactorily facilitate density forecast evaluation and comparison. Section 5 concludes.

2. The statistical framework

2.1. Calibration

We consider probabilistic forecasts in the form of predictive cumulative distribution functions (CDFs) $F_t,\ t=1,2,\ldots.$ These may be based on statistical models, supplemented by expert judgment. The outcome X_t is a random variable with distribution G_t, which represents the true data-generating process. If $F_t = G_t$ for all *t*, GBR speak of the 'ideal' forecaster; in economic forecasting this is commonly referred to as a 'rational' or 'full-information' forecast.

In making forecasts for the future, Dawid's prequential forecaster, at any time *t*, with the values $\mathbf{x}^{(t)}$ of the sequence $\mathbf{X}^{(t)} = (X_1, X_2, \ldots, X_t)$ to hand, issues a forecast distribution F_{t+1} for the next observation X_{t+1}. As noted above, the standard tool for assessing forecast performance on the basis of the forecast-observation pairs is the sequence of probability integral transform (PIT) values

$$p_t = F_t(x_t).$$

If F_t coincides with G_t, $t=1,2,\ldots$, then the p_ts are independent uniform $U[0,1]$ variables. An advantage of basing forecast evaluation on the PIT values is that it is not necessary to specify G_t, real or hypothesised. Uniformity is often assessed in an exploratory manner, by

inspection of histograms of PIT values, for example, while formal tests of goodness-of-fit are also available, as are tests of independence, described below. Diebold, Gunther and Tay (1998) introduce these ideas to the econometrics literature and provide a full proof of the iid$U$[0,1] result.

GBR define *probabilistic calibration* of the sequence F_t relative to the sequence G_t as the condition

$$\frac{1}{T}\sum_{t=1}^{T} G_t\left(F_t^{-1}(p)\right) \to p \quad \text{for all } p \in (0,1). \tag{1}$$

Their theorem 2 (2007, p.252) shows that probabilistic calibration is equivalent to the uniformity of the PIT values. Intuitively, and dropping time subscripts for convenience, given a CDF $G(x)$ and the transformation $p = F(x)$, the standard change-of-variable approach gives the CDF $H(p)$, say, as the expression inside the summation in equation (1): if $H(p) = p$ then p has a uniform distribution. Condition (1) is a convenient device for checking probabilistic calibration in circumstances where G_t is known, as in simulation experiments or theoretical exercises which require the data-generating process to be specified. We note, however, that neither this definition nor any other discussion of the theoretical framework in Section 2 of GBR's article makes any reference to the independence component of the proposition discussed in the preceding paragraph, although occasional references appear elsewhere in the article.

2.2. Statistical tests

Smith (1985) describes diagnostic checks that can be applied to a range of forecasting models, based on the PIT values p_t or on the values given by their inverse normal transformation, $z_t = \Phi^{-1}(p_t)$, where $\Phi(\cdot)$ is the standard normal distribution function. If p_t is iid$U$(0,1), then z_t is iid$N$(0,1). The advantages of this second transformation are that there are more tests available for normality, it is easier to test autocorrelation under normality than uniformity, and the normal likelihood can be used to construct likelihood ratio tests. For a density forecast explicitly based on the normal distribution, the double transformation returns z_t as the standardised value of the outcome x_t, which could be calculated directly.

Formal tests of goodness-of-fit include the classical Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests for uniformity, together with the Doornik-Hansen (DH) test for normality (Doornik and Hansen, 1994). These are all based on random sampling assumptions, and there are no general results about their performance under autocorrelation.

Test of independence can be based on the p_t or z_t series, as noted above. For the PIT series a common choice is the Ljung-Box (LB) test, based on autocorrelation coefficients up to a specified maximum lag, approximately distributed as chi-square with that specified number of degrees of freedom under the null. For their inverse normal transforms a widely used parametric test is due to Berkowitz (2001). Under a maintained hypothesis of normality, this tests jointly for correct mean and variance ('goodness-of-fit'), and independence. It is a likelihood ratio test, approximately distributed as chi-square with three degrees of freedom. For the independence component of the joint null hypothesis the alternative is that z_t follows a first-order autoregressive process

2.3. Scoring rules and distance measures

Scoring rules evaluate the quality of density forecasts by assigning a numerical score based on the forecast and the subsequent realisation of the variable. A popular choice, originally proposed by Good (1952), is the logarithmic score

$$\log S_j(x_t) = \log f_{jt}(x_t)$$

for forecast density f_{jt}. To a Bayesian the score is the predictive likelihood, and if two forecasts are being compared, the log Bayes factor is the difference in their logarithmic scores. If one of the forecasts is the correct conditional density g_t, the 'ideal' forecast, then the expected difference in their logarithmic scores is the Kullback-Leibler information criterion (KLIC) or distance measure

$$\mathrm{KLIC}_t = E\left\{\log g_t(x_t) - \log f_{jt}(x_t)\right\} = E\left\{d_t(x_t)\right\},$$

say, where the expectation is taken in the correct distribution. Interpreting the difference in log scores, $d_t(x_t)$, as a density forecast error, the KLIC can be interpreted as a mean error in a similar manner to the use of the mean error or bias in point forecast evaluation.

To develop a KLIC-based test for density forecast evaluation, Bao, Lee and Saltoglu (2004, 2007) and Mitchell and Hall (2005) replace *E* by a sample average and use

transformed variables z_t. As above, the attraction of using transformed variables z_t is that it is not necessary to know g_t, but simply that, under the null that g_t is correct, the distribution of z_t is standard normal. Or equivalently, using the PITs p_t, the null distribution is uniform and its log density is zero.

For two density forecasts f_{jt} and f_{kt}, these authors also develop a test of equal predictive accuracy based on their KLIC difference. Again replacing *E* by a sample average, but without transforming the data, a likelihood ratio test of equal forecast performance is based on the sample average of

$$\log f_{jt}(x_t) - \log f_{kt}(x_t) .$$

Amisano and Giacomini (2007) develop the same test by starting from the logarithmic score as a measure of forecast performance.

Finally we note that some simple relations are available when density forecasts are based on normal distributions, as in the examples in the next two sections. The expected logarithmic score of the correct conditional density is a simple function of its forecast variance ('sharpness'), namely

$$E_g\left\{\log g(x)\right\} = -\tfrac{1}{2}\log\left(2\pi\sigma_g^2\right) - \tfrac{1}{2} .$$

With a competing forecast $f(x)$ we obtain, subscripting parameters appropriately,

$$E_g\left\{\log g(x) - \log f(x)\right\} = -\tfrac{1}{2} - \tfrac{1}{2}\log\left(\sigma_g^2 \big/ \sigma_f^2\right) + \tfrac{1}{2}\frac{\sigma_g^2}{\sigma_f^2} + \frac{\left(\mu_g - \mu_f\right)^2}{2\sigma_f^2} .$$

The KLIC has a minimum at zero: the sum of the first three terms on the right-hand side is non-negative, as is the fourth term. Thus a positive KLIC may result from departures in mean and/or variance in either direction, and additional investigation is needed to discover the direction of any departure. The competing forecast may be too sharp or not sharp enough, but the sharpness criterion, being 'subject to calibration', would not arise if the forecast was already rejected by a KLIC-based test.

3. GBR's example

The scenario for the simulation study is that, each period, nature draws a standard normal random number μ_t and specifies the data-generating distribution $G_t = N(\mu_t, 1)$. Four competing forecasts are constructed. The ideal forecaster conditions on the current state and issues the forecast $F_t = G_t$. The 'climatological' forecaster, having historical experience in mind, takes the unconditional distribution $F_t = N(0, 2)$ as their probabilistic forecast. The remaining two forecasts are based on mixtures of models, motivated by an example of Hamill (2001). Hamill's forecaster is a master forecaster who assigns the forecasting problem with equal probability to any of three student forecasters, each of whom is forecasting incorrectly: one has a negative bias, one has a positive bias, and the third has excessive variability. Thus the forecast distribution is $N\left(\mu_t + \delta_t, \sigma_t^2\right)$, where $\left(\delta_t, \sigma_t^2\right) = (0.5,1),\ (-0.5,1)$ or $(0,1.69)$, each with probability one-third. Similarly GBR's 'unfocused' forecaster observes the current state but adds a distributional bias as a mixture component, giving the forecast distribution $0.5\{N(\mu_t,1) + N(\mu_t + \tau_t, 1)\}$ where $\tau_t = \pm 1$, each with probability one-half. With 10,000 random draws of x_t from G_t, GBR obtain the PIT histograms for the four forecasters shown in Fig. 1 (reproduced from the original).

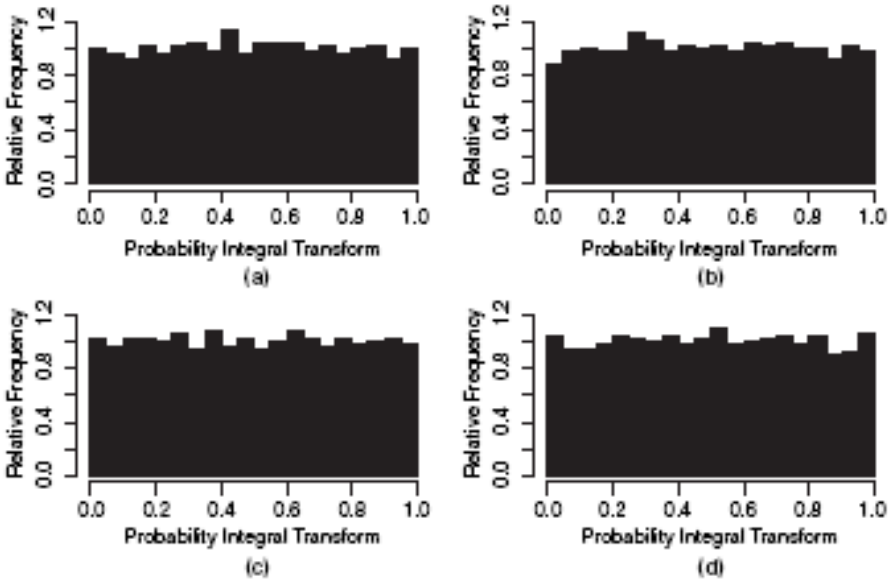

Fig. 1. PIT histograms for (a) the ideal forecaster, (b) the climatological forecaster, (c) the unfocused forecaster and (d) Hamill's forecaster

The four PIT histograms are 'essentially uniform', which 'is a disconcerting result' (2007, p. 245), because these PIT histograms cannot distinguish the ideal from the competing forecasts.

The climatological or unconditional forecaster is the first of the ideal forecaster's indistinguishable competitors. Its distribution is correctly stated, but in typical time series forecasting problems time dependence gives simple criteria for distinguishing between conditional and unconditional forecasts. Autocorrelation in the point forecast errors or density forecast PITs can be expected from an unconditional forecast, denying the independence component of Dawid's calibration condition. However the GBR example is concerned with forecasting white noise. It has the same structure as an example given by Granger (1983), in a paper entitled 'Forecasting white noise', although his formulation takes an explicit time-series forecasting perspective: 'if $x_t = y_{t-1} + e_t$ where y_t, e_t are independent, pure white noise series, then if y_t is observable, x_t will be pure white noise but forecastable ... Thus, x_t is not forecastable just from its own past but becomes forecastable when past values of y_t are added to the information set' (1983, p.308). From a practical forecasting perspective, in discrete time, the assumption in GBR's example that the state variable μ_t is observable at time *t* but the outcome x_t is not has an economic counterpart in which state variables such as tax rates are known in advance but outcomes are known only after some data delivery delay, hence the interest in 'nowcasting'. However, forecasting a white noise process is scarcely a representative example in time-series forecasting, and to better motivate a fuller discussion of relevant criteria we introduce time dependence in a second example in the next section.

The remaining forecasts are based on model mixtures or switching models, in which the forecast issued is one of two (the unfocused case) or three (Hamill's) possible forecasts, none of which have the correct distribution, chosen at random. This is in direct contrast to the forecast combination literature, which since the seminal article by Bates and Granger (1969) has considered situations in which multiple forecasts of the same variable are available at each point in time. Several competing models might be in use simultaneously, several individuals might provide their different forecasts in response to a survey, and so on; Timmermann (2006) provides a recent survey of research on forecast combinations. If we assume, in contrast to GBR's approach, that the two or three component forecasts in each of

these cases are all available at each point in time, and are combined or pooled in accordance with this literature, then the resulting finite mixture distribution forecasts have non-uniform PITs and can be readily distinguished from the ideal forecast.

To consider the more formal evaluation procedures discussed above, we reproduce GBR's experiment, constructing 500 replications of an artificial sample of size 150. For the statistical tests described in Section 2.2 the results are exactly in accordance with the informal appraisal above. We find that the three tests of fit and the two tests of independence all have rejection frequencies close to the nominal size of the tests, which we set at the usual 5% level: the 'disconcerting result' continues to apply. However we find that the KLIC-based procedures discussed in Section 2.3 are able to distinguish the ideal forecast from its competitors. In 500 replications the KLIC-based test always rejects the unconditional forecaster, while the rejection frequencies for the unfocused forecaster and Hamill's forecaster are respectively 88% and 82%. The question of sharpness is redundant, although it might be argued that it was present all along, in view of the expressions presented at the end of Section 2.

In sum, its use of a white noise data generating process and its unusual approach to forecast combination, together with the shortage of formal evaluation procedures, make GBR's example an unrealistic guide to developments in this area.

4. Forecasting an autoregressive process

4.1. The ideal forecast and five competing forecasts

Consider the Gaussian second-order autoregressive data generating process

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \;\; \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right).$$

The true or 'ideal' forecast distribution of Y_t given observations y_{t-1} and y_{t-2} and knowledge of parameter values is

$$G_t = N\left(\phi_1 y_{t-1} + \phi_2 y_{t-2}, \sigma_\varepsilon^2\right).$$

The 'climatological' or unconditional probability forecast is

$$F_{Ut} = N\left(0, \sigma_y^2\right)$$

where $\sigma_\varepsilon^2 = (1-\phi_1\rho_1-\phi_2\rho_2)\sigma_y^2$ and ρ_i, $i=1,2$, are autocorrelation coefficients:

$$\rho_1 = \phi_1/(1-\phi_2), \quad \rho_2 = \phi_1\rho_1 + \phi_2.$$

We consider a variant forecaster who assumes that the data are generated by a first-order autoregression and issues the forecast

$$F_{1t} = N\left(\rho_1 y_{t-1}, \sigma_1^2\right)$$

while, with the same assumption, a second variant is subject to a one-period data delay, so the forecast is

$$F_{2t} = N\left(\rho_2 y_{t-2}, \sigma_2^2\right),$$

where $\sigma_1^2 = \left(1-\rho_1^2\right)\sigma_y^2$ and $\sigma_2^2 = \left(1-\rho_2^2\right)\sigma_y^2$. We assume that these forecasters use least-squares regression of y_t on its relevant lagged value to estimate the required coefficient and the associated residual variance, but as above we neglect parameter estimation error and use the corresponding 'true' values. In our tables we label them AR1 and AR2 respectively.

Next is an 'averaging' forecaster who knows that forecast combination can often be of benefit and so constructs the equally-weighted combined forecast

$$F_{Ct} = 0.5N\left(\rho_1 y_{t-1}, \sigma_1^2\right) + 0.5N\left(\rho_2 y_{t-2}, \sigma_2^2\right),$$

which is an example of a finite mixture distribution (Everitt and Hand, 1981). The composite information set for the combined density forecast is identical to the information set of the true forecast density: both contain the same two observations. However the combined forecast uses the information inefficiently, relative to the true forecast. It yields, despite the fact that the true distribution is Gaussian, a mixture normal density forecast.

Finally, in contrast with the combined forecast we follow GBR's example and consider an 'unfocused' forecaster who uses a mixture of models, switching between them at random. As in their example, each model adds distributional bias to the ideal forecast, thus

$$F_{Mt} = 0.5\left\{G_t + N\left(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \tau_t, \sigma_\varepsilon^2\right)\right\},$$

where τ_t is either 1 or –1, each with probability one-half.

The performance of these six forecasts is assessed in a simulation study, using the evaluation criteria discussed above. To assess the effect of time dependence on the performance of these criteria we consider four pairs of values of the autoregressive parameters ϕ_1 and ϕ_2, as shown in Table 1. Each delivers a stationary process, with differing degrees of autocorrelation, also as shown in Table 1. Case (2) exhibits rather less persistence than is observed in the variables for which central banks often publish density forecasts, namely inflation and GDP growth. The structure of case (3) is such that the AR1 forecast F_{1t} coincides with the unconditional forecast F_{Ut}, while the AR2 forecast F_{2t} coincides with the ideal forecast G_t, thus the combined forecast F_{Ct} is a combination of the correct conditional and unconditional forecasts in this case. We report results based on 500 replications and a sample size $T = 150$, which is typical of applications in macroeconomics.

4.2. PIT histograms

We first present histograms of PIT values, to allow an informal assessment of their uniformity and hence of probabilistic calibration in GBR's sense. We note that the assumed distributions of the density forecasts, in respect of their normality and their first two moments conditional on the information on which they are based, are correct for all but the combination forecast, while the unfocused forecaster's biases are expected to be offsetting. The results presented in Fig. 2 are then completely as expected.

The PIT histograms in all columns of Fig. 2 but the fifth are 'essentially uniform': they cannot distinguish the ideal forecast from these competitors. Its fourth competitor, the combination forecast, despite being a combination of densities which each deliver uniform PITs, has too great a variance in all cases, hence all four PIT histograms in the fifth column of Fig. 2 have a humped shape. This is most striking in case (3) where, of the two forecasts being combined, the AR1 forecast, which here coincides with the unconditional forecast, has an error variance ten times greater than that of the AR2 or ideal forecast.

4.3. Statistical tests

We first consider the goodness-of-fit tests discussed in Section 2.2. Table 2 reports the rejection percentages for the KS and AD tests of uniformity of the PITs and the DH test of normality of their inverse normal transforms, all at the nominal 5% level, for each of the six density forecasts. For the KS and AD tests we use simulated critical values for the sample

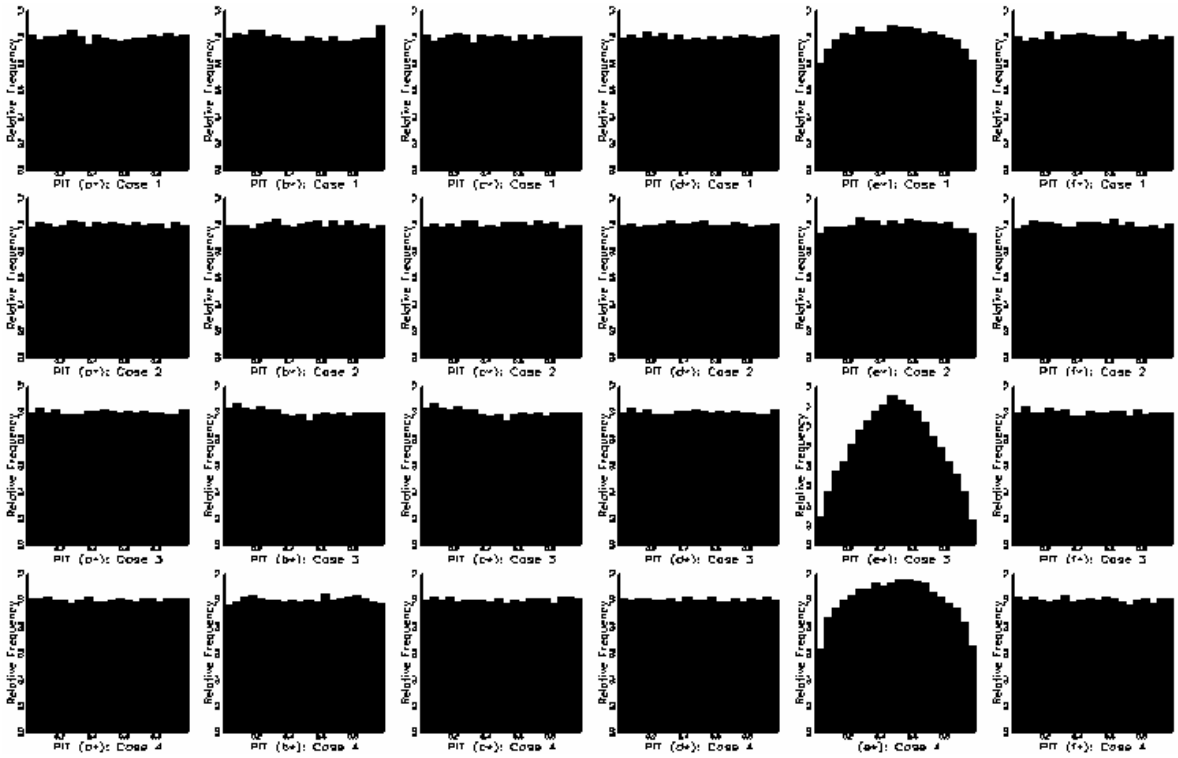

Fig. 2. PIT histograms in Cases 1-4 for (a*) ideal, (b*) climatological, (c*) AR1, (d*) AR2, (e*) combination and (f*) unfocused forecasters

size of 150, while the DH test statistic is approximately distributed as chi-square with two degrees of freedom under the null.

The results show that the goodness-of-fit tests tend not to reject any of the conditional forecasts. The rejection rates for the ideal forecasts, which have white noise errors, are not significantly different from the nominal 5% level. Autocorrelation clearly affects the performance of the tests for the two variant forecasts and their combination, but, in general, the rejection rate is not greatly increased. The unconditional or 'climatological' forecast has the greatest error autocorrelation, and this is associated with a substantial increase in the rejection rates of the goodness-of-fit tests. In case (3) the AR1 forecast and the unconditional forecast coincide, and the high rejection rate in this case also spills over to the combined forecast. In other cases these tests suggest that the combined forecast's normal mixture distribution appears not to deviate too much from normality. Nevertheless, given its non-normal distribution, the results in the fifth row suggest that the AD test has a slight advantage in power over the KS test, which is consistent with results obtained by Noceti, Smith and Hodges (2003) in the white noise case.

Turning to tests of independence, we consider the Ljung-Box test based on autocorrelation coefficients of the PIT series up to lag four, and the likelihood ratio test of Berkowitz (2001) based on the z_t series, as discussed in Section 2.2. In the present experiment the first four forecasts have mean and variance of z_t equal to (0,1), so here the test is in effect a test of the first-order autocorrelation coefficient of the point forecast errors.

The results in Table 3 show that adding a test of independence to the evaluation tool-kit immediately enables us to distinguish the ideal forecast from all the competing forecasts except the 'unfocused' mixture of models. The rejection rates for the ideal forecasts are close to the nominal size of the (asymptotic) tests, and adding a random bias does not induce autocorrelation, as seen in the last row of the table. For the remaining forecasts the tests have good power: in case (1), a relatively persistent series, there are no Type 2 errors in our 500 replications for any of the competing forecasts. This is also true of the unconditional forecasts in cases (3) and (4). Whereas Figure 2 might be thought to represent a 'disconcerting result' since it does not distinguish the ideal forecast from four of its competitors, we see that considering both components of the calibration requirement, namely independence and uniformity of the PITs, delivers the desired discrimination in three of these cases.

4.4. Scoring rules and distance measures

Average logarithmic scores and hence KLICs can be calculated from simulation results, and in the present example we can also calculate expected logarithmic scores for four of our forecasts using expressions akin to those presented at the end of Section 2.3. For the two forecasts with mixture components, the corresponding expectation can be obtained by numerical integration. Table 4 then reports the KLIC value together with the average logarithmic score for each forecast (both multiplied by 100); with the present sample size and number of replications the simulation-based average score scarcely deviates from the expected score calculated analytically.

Since the KLIC of a given forecast is the difference between its logarithmic score and that of the ideal forecast, the two criteria in Table 4 rank the forecasts identically. The differences reflect the value of the information used by each forecast in each case, except that the cost of the unfocused forecaster's addition of random biases to the ideal forecast is not

affected by the persistence of the series. The unconditional or climatological forecast uses no information from past data and is ranked last except in case (2), where the data are least persistent. The AR1 and AR2 forecasts use only a single past observation and occupy intermediate ranks, as does their equally-weighted combination. In each of cases (2) and (4) the two AR forecasts perform rather similarly, and their equally-weighted combination achieves an improvement. On the other hand in cases (1) and (3) the two AR forecasts have rather different scores so the optimal weights for a combined forecast are rather different from equality, and the equally-weighted combination takes an intermediate value.

Finally we turn to the KLIC-based test discussed in Section 2.3. Tests are undertaken for each of the competing forecasts against the ideal forecast, and rejection percentages are reported in Table 5. (We remember that in case (3) the AR2 forecast coincides with the ideal forecast.) The rejection rate for the unfocused forecaster is close to that obtained in GBR's example, while in several other cells of the table it reaches 100%. Again the power of the test is lowest in case (2), where time dependence is below the levels commonly observed and these time-series forecasts are relatively similar to one another.

In this example established criteria provided an adequate basis for distinguishing between competing forecasts, and there is no need for an additional criterion, such as their sharpness, to evaluate density forecasts. As noted above, there is a sense in which it is already subsumed in existing information-based methods.

5. Conclusion

Density forecasts are receiving increasing attention in time-series forecasting. They are becoming increasingly prevalent, which can only be welcomed, and methods of assessment are becoming increasingly available. This paper reviews some currently-available procedures for density forecast evaluation, in the light of a recent proposal by Gneiting, Balabdaoui and Raftery (2007) to add a 'sharpness' criterion to the existing tool-kit.

Since Dawid (1984), the basic foundation of density forecast evaluation, on which many subsequent developments rest, has been a two-component calibration criterion, requiring uniformity and independence of the PITs. In the example which motivates GBR's

proposal, the first component of these two cannot distinguish between the ideal forecast and its competitors, and the second is irrelevant, because their example has no time dimension. This is a surprising omission in an article that opens with the statement that 'A major human desire is to make forecasts for the future', and it might in turn be said to make their example irrelevant. An artificial construct in which there is no connection between present and future is an insecure foundation for a claim about the adequacy or otherwise of existing forecast evaluation methods. Moreover their indistinguishable competing forecasts are constructed using an approach to forecast combination which is at variance with the existing forecast combination literature; nevertheless we show that information-based methods are able to supply the required discrimination. In our alternative example, in which the variable we wish to forecast exhibits typical time dependence, we show that the two-component calibration criterion and information-based methods remain fit for purpose, and there is no call for a subsidiary criterion of sharpness.

References

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, 25, 177-190.

Bao, Y., Lee, T-H. and Saltoglu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, 26, 203-225. First circulated as 'A test for density forecast comparison with applications to risk management', University of California, Riverside, 2004.

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20, 451-468.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, 19, 465-474.

Dawid, A.P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society* A, 147, 278-290.

Diebold, F.X., Gunther, T.A. and Tay, A.S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39, 863-883.

Doornik, J A. and Hansen, H. (1994). An omnibus test for univariate and multivariate normality. Discussion Paper, Nuffield College, Oxford.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society* B, 69, 243-268.

Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society* B, 14, 107-114.

Granger, C.W.J. (1983). Forecasting white noise. In *Applied Time Series Analysis of Economic Data* (A. Zellner, ed.), pp.308-314. Economic Research Report ER-5, US Bureau of the Census.

Hamill, T.M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550-560.

Mitchell, J. and Hall, S.G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation. *Oxford Bulletin of Economics and Statisics*, 67, 995-1033.

Noceti, P., Smith, J. and Hodges, S. (2003). An evaluation of tests of distributional forecasts. *Journal of Forecasting*, 22, 447-455.

Smith, J.Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, 4, 283-291.

Tay, A.S. and Wallis, K.F. (2000). Density forecasting: a survey. *Journal of Forecasting*, 19, 235-254. Reprinted in *A Companion to Economic Forecasting* (M.P. Clements and D.F. Hendry, eds), pp.45-68. Oxford: Blackwell, 2002.

Timmermann, A. (2006). Forecast combinations. In *Handbook of Economic Forecasting* (G. Elliott, C.W.J. Granger and A. Timmermann, eds), pp.135-196. Amsterdam: North-Holland.

Tables

Table 1. Simulation design*

	Parameter		Autocorrelation	
	ϕ_1	ϕ_2	ρ_1	ρ_2
Case (1)	1.5	–0.6	0.94	0.80
Case (2)	0.15	0.2	0.19	0.23
Case (3)	0	0.95	0	0.95
Case (4)	–0.5	0.3	–0.71	0.66

* σ_ε^2 =1 in all cases

Table 2. Goodness-of-fit tests: rejection percentages at nominal 5% level*

Forecast	Case (1)			Case (2)			Case (3)			Case (4)		
	KS	AD	DH	KS	AD	DH	KS	AD	DH	KS	AD	DH
Ideal	4.6	4.4	6.4	4.0	4.4	6.2	4.2	4.2	5.4	6.0	5.2	6.0
Climt	60	66	43	14	18	6.0	86	89	56	4.4	8.4	5.0
AR1	0.8	1.0	6.4	9.4	8.8	6.6	86	89	56	13	16	5.6
AR2	6.6	8.6	12	7.8	6.8	5.6	4.2	4.2	5.4	0.2	0	3.0
Combo	5.6	6.0	8.2	7.8	8.0	6.0	93	97	11	6.8	7.2	7.8
Unfocus	4.0	5.2	6.4	5.2	4.8	4.6	6.6	5.8	6.2	5.2	5.0	5.4

*Monte Carlo standard error $\approx 1\%$ under H_0. KS is the Kolmogorov-Smirnov test, AD the Anderson-Darling test and DH the Doornik-Hansen test.

Table 3. Tests of independence: error autocorrelations and rejection percentages

Forecast	Case (1)			Case (2)			Case (3)			Case (4)		
	$\rho_1(e)$	LB	Bk	$\rho_1(e)$	LB	Bk	$\rho_2(e)$*	LB	Bk	$\rho_1(e)$	LB	Bk
Ideal	0	4.4	4.2	0	3.8	4.6	0	6.2	5.6	0	5.2	3.4
Climt	.94	100	100	.19	68	53	.95	100	99	–.71	100	100
AR1	.56	100	100	–.04	43	17	.95	100	99	.21	78	62
AR2	.77	100	100	.15	24	30	0	6.2	5.6	–.35	99	97
Combo	.73	100	100	.06	16	14	.80	98	100	–.16	35	62
Unfocus	–.01	4.4	3.8	–.01	5.0	5.4	–.01	5.0	5.0	–.01	4.6	4.2

* $\rho_1(e) = 0$ for all forecasts in Case (3) except the unfocused forecast, where $\rho_1(e)$ is repeated. LB is the Ljung-Box test and Bk the likelihood ratio test of Berkowitz (2001)

Table 4. Additional evaluation criteria: KLIC and (negative) average logarithmic score

Forecast	Case (1)		Case (2)		Case (3)		Case (4)	
	KLIC	–log*S*	KLIC	–log*S*	KLIC	–log*S*	KLIC	–log*S*
Ideal	0	142	0	142	0	142	0	142
Climt	128	270	3.83	145	117	258	39.6	182
AR1	22	164	2.04	144	117	258	4.8	147
AR2	75	217	1.16	143	0	142	11.9	154
Combo	43	185	0.71	142	35	177	3.3	145
Unfocus	11	153	11.0	153	11	153	11.0	153

Table 5. Tests of KLIC differences vs. the ideal forecaster: rejection percentages

Forecast	Case (1)	Case (2)	Case (3)	Case (4)
Climt	100	39	100	100
AR1	98	25	100	46
AR2	100	15	n.a.	93
Combo	100	10	100	55
Unfocus	87	87	87	90