EVALUATING DENSITY FORECASTS: FORECAST COMBINATIONS, MODEL MIXTURES, CALIBRATION AND SHARPNESS

JAMES MITCHELL[a] AND KENNETH F. WALLIS[b]*

[a] *National Institute of Economic and Social Research, London, UK*
[b] *Department of Economics, University of Warwick, Coventry, UK*

SUMMARY

This paper reviews current density forecast evaluation procedures, and considers a proposal that such procedures be augmented by an assessment of 'sharpness'. This was motivated by an example in which some standard evaluation procedures using probability integral transforms cannot distinguish the ideal forecast from several competing forecasts. We show that this example has some unrealistic features from a time series forecasting perspective, and so provides insecure foundations for the argument that existing calibration procedures are inadequate in practice. Our alternative, more realistic example shows how relevant statistical methods, including information-based methods, provide the required discrimination between competing forecasts. We introduce a new test of density forecast efficiency. Copyright © 2010 John Wiley & Sons, Ltd.

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. Assessment then 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. Several ways of addressing this question have been developed in the intervening years. More general density forecast evaluation and comparison procedures now include information-based methods.

This paper reviews current density forecast evaluation procedures, in the light of Gneiting *et al.*'s (2007) recommendation that such procedures be augmented by an assessment of 'sharpness'.

---

* Correspondence to: Kenneth F. Wallis, Department of Economics, University of Warwick, Coventry CV4 7AL, UK. E-mail: K.F.Wallis@warwick.ac.uk

They propose the paradigm of maximising the sharpness of the predictive distributions subject to calibration for the evaluation of forecasts. By sharpness they mean the concentration or precision of the predictive distributions, which is a property of the forecasts only, although the condition of calibration remains a property of the forecast–observation pairs. They motivate their proposal 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 these forecasts. Since one of them is the 'ideal' or correct forecast, 'this is a disconcerting result' (Gneiting *et al.*, 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. These shortcomings are elaborated below, where we show that existing evaluation procedures can overcome the 'disconcerting result'. We then provide a more realistic example in which several competing forecasts produce uniform probability integral transforms, yet again 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. It is interesting that one of our competing forecasts cannot produce uniform probability integral transforms, despite being a combination of two forecasts that can, which seems to be true more generally. We see that existing information-based procedures already subsume the sharpness/concentration/precision criterion to some extent, and it requires no additional emphasis. An extension to these procedures, to test the efficiency of density forecasts, is introduced.

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 contains our reappraisal of the example of Gneiting *et al.* (2007), hereafter GBR. Section 4 presents a second example, in which we show that 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

Probabilistic forecasts are represented as predictive cumulative distribution functions (CDFs) F_t or densities 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 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.

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 (GBR, 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 this definition of probabilistic calibration makes no reference to the independence component of the proposition discussed in the preceding paragraph. To make the distinction clear, we refer to the two-component condition as posed by Dawid—uniformity and independence of the PITs—as *complete calibration*.

Diebold *et al*. (1998) introduce these ideas to the econometrics literature and provide a full proof of the i.i.d. $U[0,1]$ result. In this literature it is usually important to pay attention to the information set on which a forecast is based, its content and its timing. Denoting the set of all information relevant to the determination of the outcome X_t, available at the time the forecast was made, as Ω_t, we write the 'ideal' forecast or correct conditional distribution as $G_t(x_t|\Omega_t)$; in economic forecasting this is commonly referred to as the 'rational' or 'efficient' forecast. A practical forecast $F_t(x_t|W_t)$, based on a different, possibly incomplete information set, say W_t, might have a different functional form, representing a different distribution, with different moments. We denote the correct distribution conditional on that given information set as $G_t^*(x_t|W_t)$, retaining the letter G to represent its ideal or correct nature but with an asterisk to indicate that this is with respect to a different information set. Then we observe that if a practical forecast $F_t(x_t|W_t)$ coincides with the correct conditional distribution $G_t^*(x_t|W_t)$ it satisfies probabilistic calibration—it has uniform PITs—but not necessarily complete calibration (see, for example, Corradi and Swanson, 2006). Analogously, a point forecast based on incomplete information may have zero-mean errors and so be unbiased, but it may not necessarily be efficient in a mean squared forecast error sense.

2.2. Statistical Tests of Calibration

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 $\Phi(.)$ is the standard normal distribution function. If p_t is i.i.d. $U(0,1)$, then z_t is i.i.d. $N(0,1)$. 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: if the mean of the density forecast is taken as the point forecast, then z_t is equal to the point forecast error divided by the forecast standard deviation.

Formal tests of goodness-of-fit can be based on the p_t or z_t series, as noted above. Pearson's classical chi-squared test assesses the goodness-of-fit of the PIT histogram to a uniform distribution. The empirical cumulative distribution function of the PITs can be tested for uniformity by the Kolmogorov–Smirnov (KS) test or its Anderson–Darling (AD) modification. The Doornik–Hansen (DH) test for normality of the zs uses transformed skewness and kurtosis measures (see Doornik and Hansen, 2008). These tests are all based on random sampling assumptions, and there are no general results about their performance under autocorrelation. Corradi and Swanson (2006) describe extensions to Kolmogorov type tests in the presence of dynamic misspecification and parameter estimation error in the forecasting model; Bai and Ng (2005) provide generalisations of tests based on skewness and kurtosis to dependent data.

Test of independence can likewise be based on either series. For the PIT series a common choice is the Ljung–Box (LB) test, a function of autocorrelation coefficients up to a specified maximum lag, which is 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, the joint null hypothesis of correct mean and variance ('goodness-of-fit') and independence is tested against a first-order autoregressive alternative with mean and variance possibly different from (0,1). A likelihood ratio test with three degrees of freedom is based on the estimated mean, regression coefficient and error variance of the equation

$$z_t - \mu = \rho(z_{t-1} - \mu) + \varepsilon_t, \ \ \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \tag{2}$$

A test of the significance of the estimate of ρ gives a test of the independence component alone. To generalise the distributional assumption Bao *et al.* (2007) specify a flexible alternative distribution for ε_t which nests the normal distribution, their example being a semi-parametric density function, and include the additional restrictions that reduce it to normality among the hypotheses under test.

2.3. Scoring Rules, Distance Measures and Sharpness

Scoring rules evaluate the quality of probability forecasts by assigning a numerical score based on the forecast and the subsequent realisation of the variable. Their literature originates in the mid 20th century, with the quadratic score for probability forecasts of a categorical variable, due to Brier (1950), and the logarithmic score for forecasts of a continuous variable, originally proposed by Good (1952). Sharpness entered the forecasting lexicon with the decomposition of the Brier score into two components by Sanders (1963), respectively measuring the 'validity' and 'sharpness' of the forecasts. Subsequent terminology equates validity with calibration or reliability, and sharpness with refinement or resolution (see Kroese and Schaafsma, 2006); both components are functions of the forecast–observation pairs, unlike 'sharpness' as redefined by GBR.

The logarithmic score for forecast density f_{jt} is defined as

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

To a Bayesian the logarithmic score is the logarithm of the predictive likelihood, and if two forecasts are being compared the log Bayes factor is the difference in their logarithmic scores: see Geweke and Amisano (2010) for a comparison of five forecasting models using predictive

likelihood functions. 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:

$$\text{KLIC}_{jt} = E\left\{\log g_t(x_t) - \log f_{jt}(x_t)\right\} = E\left\{d_{jt}(x_t)\right\}$$

say, where the expectation is taken in the correct distribution. Interpreting the difference in log scores, $d_{jt}(x_t)$, 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 *et al.* (2007) and Mitchell and Hall (2005) replace E by a sample average and use transformed variables z_t. Then the density forecast error can be written

$$d_{jt}(x_t) = \log h_{jt}(z_t) - \log \phi(z_t)$$

where $h_{jt}(.)$ is the density of z_t and $\phi(.)$ is the standard normal density. As above, the attraction of using transformed variables z_t is that it is not necessary to know g_t, but simply that for the ideal forecast the distribution of z_t is standard normal. Except in simulation experiments $h_{jt}(.)$ is unknown, but as discussed following equation (2), above, it may be parameterised so that it nests $\phi(.)$.

For two density forecasts f_{jt} and f_{kt}, these authors also develop a test of equal predictive accuracy based on their KLIC difference. A very general setting for such tests is provided by Giacomini and White (2006), whose framework encompasses point, interval and density forecasts and a wide range of loss functions. Adapting and simplifying the notation of their equations (4) and (6), the Wald-type test statistic is given as

$$T\left(T^{-1}\sum_{t=1}^{T} b_t \Delta L_t\right)' \Sigma^{-1}\left(T^{-1}\sum_{t=1}^{T} b_t \Delta L_t\right) \tag{3}$$

where b_t is a $q \times 1$ vector of test functions, ΔL_t is the difference in the loss or score of the two forecasts in period t, and Σ is an appropriate heteroscedasticity and autocorrelation consistent (HAC) estimate of the asymptotic covariance matrix. The null hypothesis is

$$E(b_t \Delta L_t) = 0, \ \ t = 1, \ldots, T$$

under which the distribution of the test statistic (3) tends to χ_q^2 as $T \to \infty$. This is proved by Giacomini and White under much more general conditions than we require below. The simplest test of equal forecast performance has $q = 1$ and $b_t = 1$, $t = 1, \ldots, T$: in this case, when using the logarithmic score, so that $\Delta L_t = \log S_j(x_t) - \log S_k(x_t)$, the test is equivalent to (the unweighted version of) the likelihood ratio test of Amisano and Giacomini (2007).

Returning to the analogy with point forecast evaluation suggested by the interpretation of $d_{jt}(x_t)$ as a density forecast error, we recall that tests of efficiency of point forecasts are often based on correlations between forecast errors and variables which might reasonably be considered to be part of the information set available to the forecaster. A significant correlation indicates that the variable in question could have been used to predict the forecast error and hence improve the forecast:

the original forecast is thus shown to be inefficient. A finding of efficiency is seldom conclusive, however, as long as the possible existence of an untested variable which might lead to rejection remains. We propose parallel tests of efficiency of density forecasts based on the orthogonality of the density forecast error to a k-dimensional information set W_t. For this purpose elements of W_t are introduced into the conditional mean of the density $h_{jt}(.)$, and their significance is tested via a likelihood ratio test. A regression as in equation (2), again with possibly more general distributional assumptions, is a convenient setting for this procedure. We note that Berkowitz (2001, p. 468) suggests that the regression equation used to implement his test could be augmented by variables that 'might indicate missing factors that should be included in the underlying forecast model', although he does not pursue this.

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. Then the expected logarithmic score of the correct conditional density is a simple function of its forecast variance (sharpness/concentration/precision), namely

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

For a competing forecast $f(x)$ we obtain the KLIC, subscripting parameters appropriately, as

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

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, via the PIT histogram, for example, is needed to discover the direction of any departure. The competing forecast may be too sharp or not sharp enough, indicated by a U-shaped or hump-shaped PIT histogram, respectively, but the sharpness criterion, being 'subject to calibration', would not arise if the forecast was already rejected by any of the tests described above.

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. We reproduce GBR's experiment, taking 10,000 random draws of x_t from G_t, and obtain the PIT histograms for the four forecasters shown in Figure 1. The four PIT histograms are 'essentially uniform', which 'is a disconcerting result' (GBR, p. 245), because these PIT histograms cannot distinguish the ideal from the competing forecasts: all four forecasts are probabilistically calibrated.

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. Economic time series commonly display persistence, and in financial examples where a white noise null hypothesis may be relevant, such as stock returns, models of persistence in volatility are commonplace. Autocorrelation in the point forecast errors or density forecast PITs can then 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
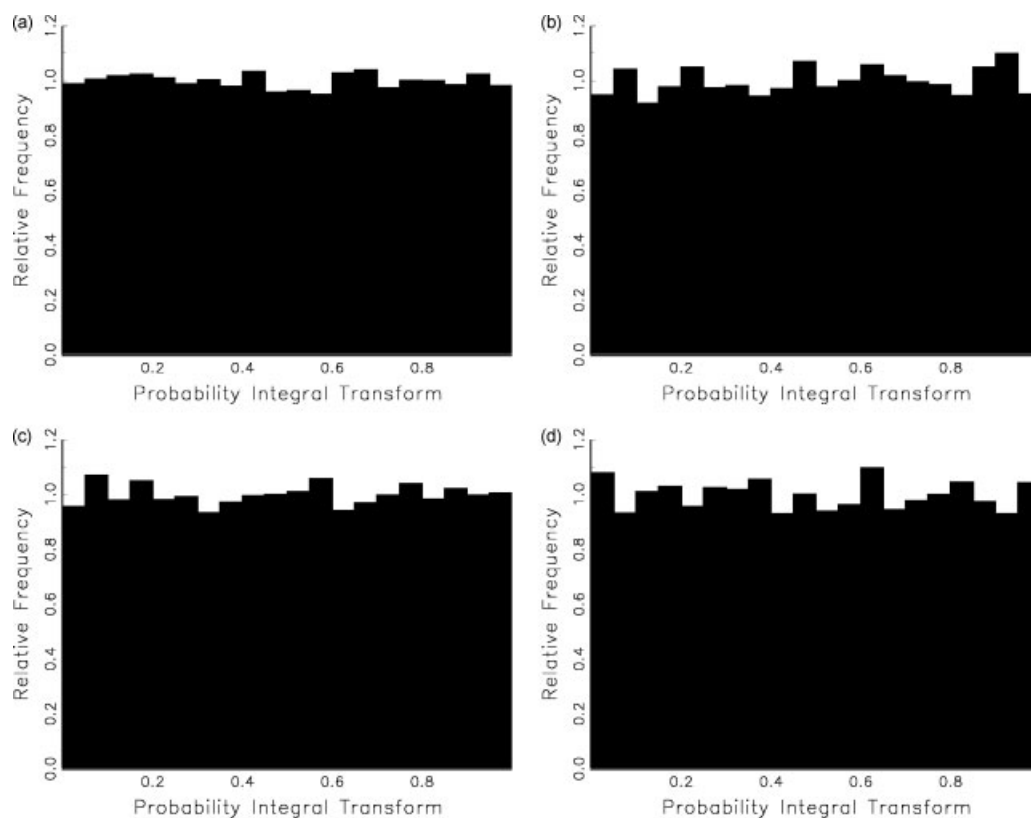

Figure 1. PIT histograms for GBR's example: (a) the ideal forecaster; (b) the climatological forecaster; (c) the unfocused forecaster; (d) Hamill's forecaster

becomes forecastable when past values of y_t are added to the information set' (Granger, 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, with many practical series displaying persistence, 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 mixtures of models or switching models, in which the forecast issued is chosen at random as one of two (the unfocused case) or three (Hamill's) possible forecasts, none of which in either case has the correct distribution. This selection mechanism is in direct contrast to the forecast combination literature, which since the seminal article by Bates and Granger (1969) has considered situations in which several available forecasts at each point in time are combined in some way to give an improved forecast. An individual forecaster might combine forecasts from several models in constructing their forecast; in response to a survey several individuals might report their different forecasts, which are then combined, and so on: Timmermann (2006) provides a recent survey of research on forecast combinations. Whereas most of the literature is concerned with point forecast combination, Wallis (2005) considers extensions to density forecast combination, for which the appropriate statistical framework is the finite mixture distribution. It is potentially confusing that the term 'mixture' or 'mixture model' is used in two different contexts, but these can be distinguished by the interpretation of the coefficients in the (linear) combination of forecasts. In the combining or pooling literature these are typically non-negative weights which sum to one, the simple average with equal weights often being recommended for point forecasts, although its effectiveness for density forecasts has been questioned (Jore *et al.*, 2010). In GBR's mixture of models the implicit coefficients are latent indicator $(0-1)$ random variables which sum to one; in regime-switching models such indicator variables often have a Markov structure. In their study of optimal weights Geweke and Amisano (2009) speak of pools rather than combinations, and in Section 6 of their paper compare the log predictive scores of the linear pool and the mixture of two forecasting models. If we similarly compare GBR's mixture of models with the linear combinations with equal weights of their two or three component forecasts, then we find that the resulting finite mixture distribution forecasts have non-uniform PITs and can be readily distinguished from the ideal forecast.

To assess sharpness, GBR subsequently report the average width of central 50% and 90% prediction intervals for the four forecasters, and their mean log scores over the sample of 10,000 replications. Both rank the ideal forecaster best, followed by Hamill's, the unfocused and the climatological forecaster. No statistical testing is undertaken of the coverage of the prediction intervals, or of the significance of the differences in log scores, for example.

We remedy this omission by utilising the evaluation procedures discussed in Section 2. We reproduce GBR's experiment and construct 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 of the PIT histograms. We find that three tests of fit and two tests of independence all have rejection frequencies close to the nominal size of the tests, which we set at the usual 5% level. Thus the 'disconcerting result' continues to apply, now with the sense that, in this white noise context, all four forecasts are completely calibrated. 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%. These results yield clear discrimination between the ideal forecast and its competitors.

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 an information set Ω_t comprising observations y_{t-1} and y_{t-2}, the model and its 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:

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

The second-order autoregression is a relatively simple model, but it gives sufficient scope for constructing some competing forecasts that can be expected to deliver uniform PITs, as in GBR's example. We seek to illustrate the consequences of the fact that probabilistic calibration holds whenever a density forecast is the correct conditional density with respect to its specific information set, and leave practical complications such as parameter estimation error and model uncertainty on one side.

Our first variant forecaster assumes that the data are generated by a first-order autoregression and so issues the forecast

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

while a further variant makes the same assumption but 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 an equally weighted combination of the two preceding forecasts:

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

The composite information set for this combined density forecast includes the same two observations as the information set of the ideal forecast. However, the combined forecast uses this information inefficiently, relative to the ideal forecast. The forecast distribution is a mixture of two normal distributions, as in Pearson's (1894) classic article that initiated the finite mixture distribution literature.

Finally, in contrast to 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, but the biases are again expected to be offsetting.

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 I. Each delivers a stationary process, with differing degrees of autocorrelation, also as shown in Table I. Although integrated series are prevalent in macroeconomics, forecasting typically focuses on transformed variables that are nearer to stationarity, such as inflation rather than the price level, and growth rather than the output level. Inflation and GDP growth are the variables for which several central banks currently publish density forecasts. Case (1) represents a relatively persistent stationary series, whereas case (2) exhibits less persistence than is observed in 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. Case (4) represents a rather unusual oscillating form. We report results based on 500 replications and a sample size $T = 150$, which is typical of applications in macroeconomics.

Table I. 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

Note: $\sigma_\varepsilon^2 = 1$ in all cases.

4.2. PIT Histograms

For initial diagnostic checking we present histograms of PIT values, to allow an informal assessment of their uniformity and hence of probabilistic calibration in GBR's sense. This is expected to hold for the first four forecasts, since each of these states the correct conditional distribution in respect of its normality and its first two moments conditional on the past data utilised by the forecaster. The results presented in Figure 2 are then completely as expected.

The PIT histograms in all columns of Figure 2 but the fifth are 'essentially uniform': they cannot distinguish the ideal forecast from these competitors. Its fourth competitor, the combination forecast, despite being an equally weighted combination of densities which each deliver uniform PITs, has too great a variance in all cases; hence too few observations lie in the tails of the forecast densities and all four PIT histograms in the fifth column of Figure 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 variance 10 times greater than that of the AR2 or ideal forecast, and its optimal weight is zero.

4.3. Statistical Tests

We first consider the goodness-of-fit tests discussed in Section 2.2. Table II reports the rejection percentages across 500 replications 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 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 in general tend to confirm the visual impressions of the PIT histograms. 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 greatest error autocorrelation occurs with the unconditional or 'climatological' forecast in cases (1) and (3), and this is associated with a substantial distortion of the distribution of the goodness-of-fit test statistics. 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 *et al.* (2003) for white noise data.

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 (standardised) point forecast errors, since these density forecasts assume a normal distribution. To aid interpretation of the performance of the tests, the theoretical first-order error autocorrelation coefficient is reported alongside the test results in Table III. In case (3) this is exactly zero for all six forecasts, and the second-order coefficient is reported instead.
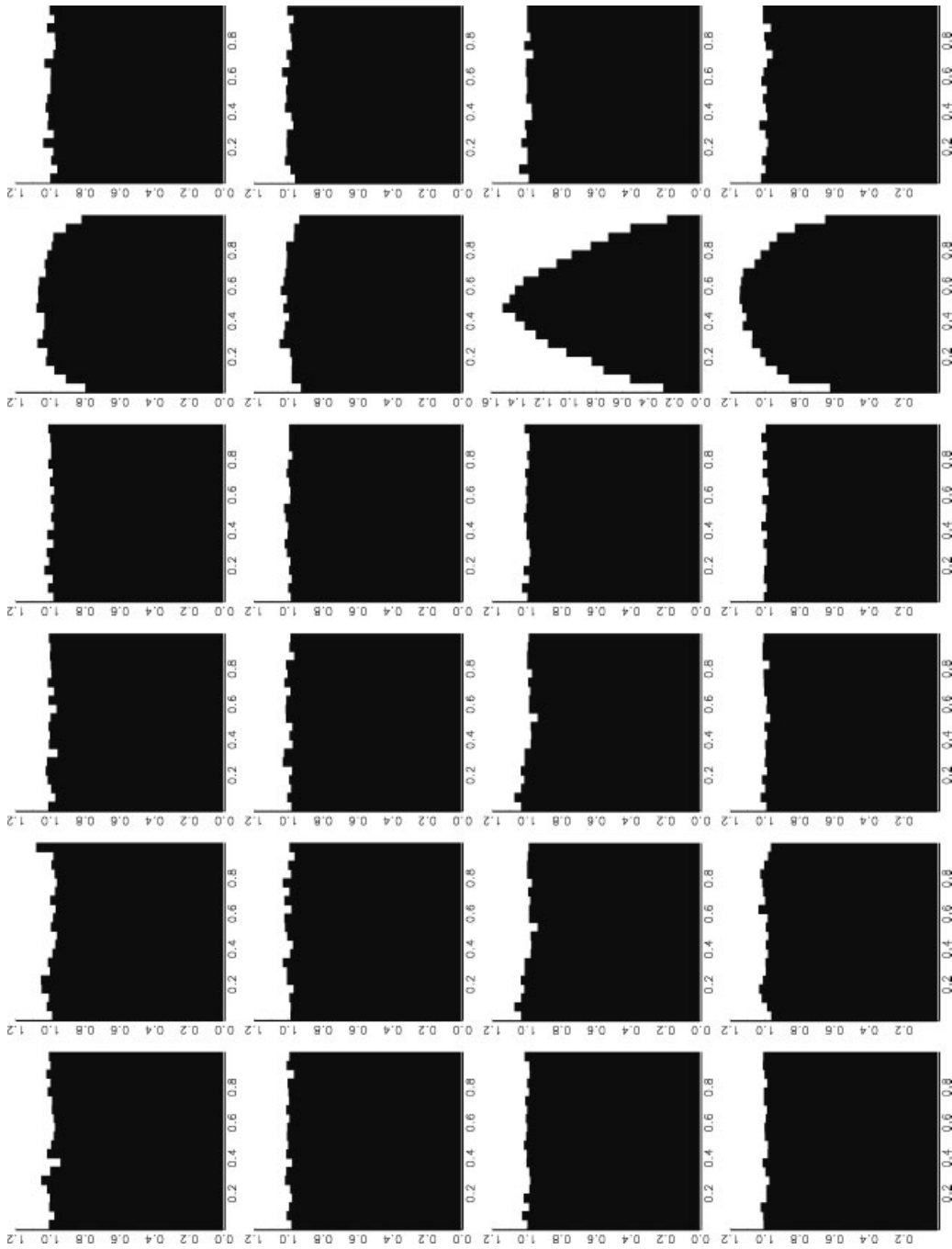
Figure 2. PIT histograms for our autoregressive example. Rows: cases (1)–(4) as defined in Table I. Columns: forecasts, respectively ideal, climatological, AR1, AR2, combination, unfocused

Table II. 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

Note: Monte Carlo standard error $\approx 1\%$ under H_0. KS, Kolmogorov–Smirnov test; AD, Anderson–Darling test; DH, Doornik–Hansen test.

Table III. Tests of independence: error autocorrelations and rejection percentages at nominal 5% level

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	0.94	100	100	0.19	68	53	0.95	100	99	−0.71	100	100
AR1	0.56	100	100	−0.04	43	17	0.95	100	99	0.21	78	62
AR2	0.77	100	100	0.15	24	30	0	6.2	5.6	−0.35	99	97
Combo	0.73	100	100	0.06	16	14	0.80	98	100	−0.16	35	62
Unfocus	0	4.4	3.8	0	5.0	5.4	0	5.0	5.0	0	4.6	4.2

Note: $\rho_j(e)$ is the jth-order autocorrelation coefficient of the point forecast errors, which is equal to the autocorrelation coefficient of z_t for the first four forecasts. In Case (3) $\rho_1(e) = 0$ for all forecasts, and $\rho_2(e)$ is reported instead. LB, Ljung–Box test: Bk, likelihood ratio test of Berkowitz (2001).

The results in Table III show that adding a test of independence to the evaluation toolkit 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 complete calibration—not only uniformity but also independence of the PITs—delivers the desired discrimination in three of these cases.

4.4. Scoring Rules, Distance Measures and KLIC-Based Tests

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 IV then reports the KLIC value together with the average logarithmic score for each forecast (both

Table IV. 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

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 expected difference between its logarithmic score and that of the ideal forecast, the two criteria in Table IV 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.

Whether the summary information in Table IV represents significant differences in forecast performance is assessed by the log score or KLIC-based test discussed in Section 2.3. We test each of the competing forecasts against the ideal forecast, and report rejection percentages across the 500 replications in Table V. (We remember that in case (3) the AR2 forecast coincides with the ideal forecast.) The rejection rate for the unfocused forecaster is seen to be close to the rejection rate for this forecaster that we obtained in the GBR example (Section 3), 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 the more typical cases, the test is shown to provide clear discrimination between the ideal forecast and its competitors, despite their uniform PIT histograms.

Finally we turn to a test of density forecast efficiency as proposed in Section 2.3. In the present example the information set is quite small, containing no extraneous variables; nevertheless

Table V. Tests of KLIC differences vs. the ideal forecaster: rejection percentages at nominal 5% level

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

Table VI. Tests of efficiency: rejection percentages at nominal 5% level

Forecast	Added regressor	Case (1)	Case (2)	Case (3)	Case (4)
Ideal	y_{t-1}	5	6	2	5
	y_{t-2}	5	6	8	5
	Both	7	5	10	4
Climt	y_{t-1}	n.a.	n.a.	n.a.	n.a.
	y_{t-2}	100	60	100	93
	Both	n.a.	n.a.	n.a.	n.a.
AR1	y_{t-1}	97	50	n.a.	67
	y_{t-2}	97	56	100	67
	Both	n.a.	n.a.	n.a.	n.a.
AR2	y_{t-1}	98	6	2	24
	y_{t-2}	100	8	8	37
	Both	100	8	10	100
Combo	y_{t-1}	99	7	48	20
	y_{t-2}	100	15	99	18
	Both	100	12	99	16

the different forecasts make different use of the available lagged values y_{t-1} and y_{t-2}, and we study the ability of the test to discriminate between more or less efficient uses of this information. To implement the test we add test variables from the information set to the regression equation (2). Since the regression equation already contains one lag of the density forecast error, which with normal densities is equal to the standardised point forecast error, it is immediately clear that, in some cases, the addition of test variable(s) will result in perfect collinearity; hence this exercise is subject to some limitations. Cases which as a result cannot be implemented are indicated by 'n.a.' in the relevant cells of Table VI. For example, for the AR1 forecast, equation (2) in effect regresses $y_t - \rho_1 y_{t-1}$ on $y_{t-1} - \rho_1 y_{t-2}$, and we can add either y_{t-1} or y_{t-2} without inducing perfect collinearity, but not both. Otherwise, Table VI reports the relative frequency, across 500 replications, with which the null hypothesis of efficiency is rejected, by virtue of the significance at the 5% level of the coefficient(s) of the added variable(s).

The results in Table VI show that the performance of the test varies with the amount of autocorrelation in the data, which provides an indication of the amount of information that is lost by inefficient use of the available variables. In case (1) the data are most highly autocorrelated and the test performs very well, with rejection rate equal to the size of the test for the ideal forecaster, and very high power to reject the null hypothesis of efficiency for all competing forecasts. This power is then reduced as the autocorrelation falls, with case (2) the weakest; nevertheless the performance of the new test is encouraging. The source of inefficiency in the unfocused forecast is the bias that is mixed into the ideal forecast, and this bias is not sensitive to the present test; the rejection rates, not reported, are very similar to those of the ideal forecast.

The data in this example exhibit varying degrees of time dependence, and we see that established criteria provided an adequate basis for distinguishing between competing forecasts. To evaluate density forecasts there is no need to place additional emphasis on their sharpness/concentration/precision, beyond the extent to 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 and considers the proposal by Gneiting *et al.* (2007) to add a 'sharpness' criterion to the existing toolkit.

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, which we term complete calibration. 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 complete calibration criterion and information-based methods are fit for purpose.

The key to this result is a time series setting in which several competing forecasts, each satisfying probabilistic calibration, can be constructed. In assuming a given model with known parameters for each competing forecast, however, several practical features are set aside. For a given model, the need to estimate its parameters means that the resulting parameter estimation error may have an influence on the performance of any hypothesis-testing procedures that are employed, and in the present context this remains a subject for further research. But in practice the model itself is uncertain. In response practical forecasters often apply personal judgment across a suite of models, each of whose forecasts may also be subject to judgmental adjustment. Formalising such methods remains a difficult task. Nevertheless the prequential principle leads to assessment on the basis of forecasts and outcomes only, and the methods considered above remain relevant to diagnostic checking and ranking of competing forecasts, whatever their source.

Two distinct strands of the general forecast comparison literature can be identified. In one, researchers have considered a wide range of forecast construction and evaluation issues by constructing several competing forecasts themselves, and going on to study their comparative forecast performance. When artificial data are employed, as in the exercises reported in this paper, the data-generating process and hence the optimal point or density forecast is known, and this provides the best or 'sharpest' forecasts. When real data are employed, the optimal forecast is not known, but the winning forecast is usually selected according to the same criteria. A second strand of the forecast comparison literature studies the real-time forecasts supplied by respondents to forecast surveys or collected by researchers from forecast publications; in economics the best-known survey is the US Survey of Professional Forecasters. In general little is known about the statistical methods and models on which these forecasts are based, and there is typically an input of individual judgment, as noted above. For interval and density forecasts, it is possible that this results in underestimation of uncertainty. Such a finding is reported by Giordani and Soderlind (2003), who study the coverage of interval forecasts of US inflation constructed from probability

forecasts reported by respondents to the SPF, and find that the intervals are too narrow, hence the 'sharpest' forecast is an inferior forecast. Whereas in theoretical exercises the best forecast is known and it is difficult to construct competing forecasts that are too 'sharp', except by subjective intervention, there are practical circumstances in which a preference for the 'sharpest' forecast is likely to lead the forecast user astray. To emphasise 'sharpness' in this way is not generally recommended.

ACKNOWLEDGEMENTS

REFERENCES

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

Bai J, Ng S. 2005. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business and Economic Statistics* **23**: 49–60.

Bao Y, Lee T-H, 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 JM, Granger CWJ. 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.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**: 1–3.

Corradi V, Swanson NR. 2006. Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics* **133**: 779–806.

Dawid AP. 1984. Statistical theory: the prequential approach. *Journal of the Royal Statistical Society A* **147**: 278–290.

Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**: 863–883.

Doornik JA, Hansen H. 2008. An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* **70**: 927–939.

Geweke J, Amisano G. 2009. *Optimal prediction pools*. Working Paper No.1017, European Central Bank, Frankfurt.

Geweke J, Amisano G. 2010. Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting* **26**: 216–230.

Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545–1578.

Giordani P, Soderlind P. 2003. Inflation forecast uncertainty. *European Economic Review* **47**: 1037–1059.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B* **69**: 243–268.

Good IJ. 1952. Rational decisions. *Journal of the Royal Statistical Society B* **14**: 107–114.

Granger CWJ. 1983. Forecasting white noise. In *Applied Time Series Analysis of Economic Data*, Zellner A (ed.). Economic Research Report ER-5, US Bureau of the Census: Washington, DC; 308–314.

Hamill TM. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**: 550–560.

Jore AS, Mitchell J, Vahey SP. 2010. Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics* **25**: 621–634.

Kroese AH, Schaafsma W. 2006. Weather forecasting: Brier score in. In *Encyclopedia of Statistical Sciences* (2nd edn), Balakrishnan N, Read CB, Vidakovic B, (eds). Wiley: New York; 9071–9072.

Mitchell J, Hall SG. 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 Statistics* **67**: 995–1033.

Noceti P, Smith J, Hodges S. 2003. An evaluation of tests of distributional forecasts. *Journal of Forecasting* **22**: 447–455.

Pearson K. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A* **185**: 71–110.

Sanders F. 1963. On subjective probability forecasting. *Journal of Applied Meteorology* **2**: 191–201.

Smith JQ. 1985. Diagnostic checks of non-standard time series models. *Journal of Forecasting* **4**: 283–291.

Tay AS, Wallis KF. 2000. Density forecasting: a survey. *Journal of Forecasting* **19**: 235–254; Reprinted in *A Companion to Economic Forecasting*, Clements MP, Hendry DF (eds). 2002. Blackwell: Oxford; 45–68.

Timmermann A. 2006. Forecast combinations. In *Handbook of Economic Forecasting*, Elliott G, Granger CWJ, Timmermann A, (eds). North-Holland: Amsterdam; 135–196.

Wallis KF. 2005. Combining density and interval forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics* **67**: 983–994.